N.N. MOISEYEV

# PROBLÈMES MATHÉMATIQUES D'ANALYSE DES SYSTÈMES

N. MOÏSSÉEV

# PROBLÈMES MATHÉMATIQUES D'ANALYSE DES SYSTÈMES

ÉDITIONS MIR • MOSCOU

# TABLE DES MATIÈRES

# AVANT-PROPOS

La discipline qui a pour nom « analyse des systèmes » est née des besoins de procéder à des études de caractère interdisciplinaire. La conception de systèmes techniques compliqués, la planification et la gestion de complexes économiques, l'analyse des situations écologiques et nombre d'autres orientations des activités technique, scientifique et économique ont nécessité une organisation des recherches qui rompît avec la tradition : concentration des efforts des scientifiques de diverses branches, unification et concordance de l'information recueillie par des recherches concrètes. Ces recherches interdisciplinaires, appelées parfois recherches complexes ou de systèmes, doivent leur succès pour une grande part aux possibilités de traitement de l'information, à la mise en œuvre de méthodes mathématiques qui ont vu le jour avec les ordinateurs, méthodes qui ont fourni non seulement un instrument, mais aussi un langage éminemment universel. Soulignons encore une fois que l'analyse des systèmes est née à l'époque des ordinateurs, et son essor dépend pour beaucoup de leurs performances actuelles et futures. Pour cette raison, le vocable « analyse des systèmes » sera pris dans cet ouvrage dans une acception assez étroite: *par analyse des systèmes on entendra l'ensemble des méthodes s'appuyant sur les ordinateurs et ayant pour objet l'étude de systèmes: techniques, économiques, écologiques, etc.* Le résultat de ces recherches sera en principe constitué par la prise d'une décision: un plan de développement d'une région, des paramètres de construction, etc. Donc, *l'analyse des systèmes est une discipline qui s'intéresse aux problèmes décisionnels lorsque la prise d'une décision implique une analyse d'une information complexe de diverse nature physique.* Par conséquent, l'analyse des systèmes et ses conceptions méthodologiques tiennent des disciplines qui s'occupent des problèmes décisionnels: la théorie de la recherche opérationnelle et la théorie générale de la commande.

Le problème de prise de décision a été et reste au cœur des préoccupations de la vie de l'homme; en effet, toute activité est en fait une suite de prises de décision. Mais dans la plupart des cas, la prise de décision, c'est-à-dire le choix d'une ligne d'action, ne se réclame d'aucune méthode scientifique: les situations sont relativement

simples et les gens s'en remettent toujours à leur expérience, à leurs habitudes traditionnelles, à leur intuition.

Mais dans les situations difficiles, le décideur a douté de son choix, d'où la nécessité de méthodes scientifiques. Ces méthodes se sont développées progressivement pour donner corps à une discipline: la théorie de la décision. A l'étape actuelle de son développement où son appareil et son outil font une large part aux ordinateurs et que le système de ses modèles s'est transformé en un système complexe et évolué, cette théorie a pris le nom d'analyse des systèmes.

S'il est difficile de dater cette discipline, une chose est cependant claire, elle remonte aux temps les plus reculés de l'histoire de l'humanité, aux premiers embryons de l'art militaire, du négoce, de la production, etc.

Mais jusqu'à un certain temps tout ce qui relevait du choix de solutions rationnelles n'était pas à proprement parler une science. Ce n'était qu'une recette de lois accumulées par la tradition ou reflétant la vision subjective de telle ou telle personne. La prise de décision a commencé à s'ériger en science seulement après l'apparition de modèles spécifiques et l'élaboration d'une méthodologie commune d'analyse des problèmes de diverse nature physique.

La formation de cette discipline remonte à la fin du XIX$^e$ — début du XX$^e$ siècle, époque à laquelle sont apparus les premiers travaux sur la théorie de la régulation et qu'en économie on a commencé à parler des solutions optimales, c'est-à-dire du temps où est apparue la notion de fonction objectif (d'utilité) et où V. Pareto a formulé le principe de compromis.

La théorie de la décision doit sa promotion, d'une part à l'évolution de l'appareil mathématique et à la mise au point de méthodes de formalisation, d'autre part, aux nouveaux problèmes posés à l'industrie, l'art militaire, l'économie.

La théorie de la décision s'est développée à pas de géant après les années cinquante grâce à une nouvelle discipline synthétique: la recherche opérationnelle, qui tire ses méthodes de la théorie de l'efficacité, la théorie des jeux et la théorie des files d'attente.

La théorie moderne de la décision s'est dotée d'un riche appareil composé de méthodes mathématiques élaborées et de puissants systèmes de calcul. Et pourtant si retentissants que soient les succès remportés par cette théorie à l'aide des toutes dernières méthodes fondées sur une description formelle des situations, l'analyse traditionnelle qui fait jouer l'expérience, l'intuition et les capacités de l'homme à associer et à percevoir les choses extra-mathématiques hors de portée encore de l'intellect artificiel, reste nécessaire et parfois même est décisive. Aussi les méthodes de l'analyse des systèmes doivent-elles obligatoirement contenir une description des procédures non formelles utilisées, faute de quoi tout jugement sur l'analyse des systèmes sera non seulement incomplet, mais tout simplement faussé. Il faut

non seulement décrire les méthodes euristiques étudiées et les raisonnements, mais aussi montrer comment ces méthodes euristiques non formelles s'inscrivent dans la théorie moderne de la décision, comment elles se métamorphosent sous l'influence des instruments forgés par cette théorie.

L'analyse des systèmes est aujourd'hui une vaste discipline synthétique composée de plusieurs sections qui sont autant de disciplines scientifiques. Toute tentative d'étudier d'une manière tant soit peu exhaustive les problèmes qui font l'objet de l'analyse des systèmes est par avance vouée à l'échec.

Dans cet ouvrage l'auteur a voulu jeter un coup d'œil global sur cette discipline: indiquer ses sources, les possibilités de ses instruments, développer les procédés formels et non formels de raisonnement susceptibles d'être combinés grâce aux systèmes de simulation sur ordinateur *).

Mais pour réussir dans cette entreprise, pour montrer comment en analyse des systèmes se combinent les principes expérimentaux, euristiques et rigoureusement mathématiques, le laconisme doit être la règle dans la description des rubriques et méthodes. Or cela signifie que l'auteur doit toucher un lecteur plus ou moins initié.

Lorsque j'ai réfléchi au contenu de cet ouvrage et à la manière de l'exposer, j'ai imaginé pour lecteur l'étudiant de fin du second cycle, se spécialisant en mathématiques appliquées, un étudiant ayant digéré un cours d'analyse de niveau universitaire (avec de solides connaissances en équations différentielles), au fait des méthodes de calcul numérique, des méthodes d'optimisation, du calcul aux variations et de bien d'autres choses. Il lui est important de voir comment ces disciplines si hétéroclites et si peu liées entre elles de prime abord s'imbriquent et commencent à être utiles pour la pratique, grâce précisément au fait que cet objectif unique agrège ces connaissances, et le mathématicien qui les a assimilées se transforme en un architecte du système, en l'un des principaux acteurs dans de vastes recherches interdisciplinaires.

Certes je voudrais mettre entre les mains de l'analyste des systèmes des recettes, l'aider aussi à trouver sa propre voie. Je répète, sa propre voie, car en étudiant un système qui est effectivement complexe, il ne suffit pas de connaître les recettes en vigueur. L'analyse de chaque système complexe est un problème unique en son genre, impliquant non seulement une riche érudition, mais aussi de l'ingéniosité et du talent. Toute recette n'est qu'un guide.

---

*) Sur le plan des idées et de la méthode, cet ouvrage est proche du travail [5] en ce sens qu'il en concrétise les principes dans le domaine des activités humaines liées aux problèmes décisionnels. La lecture de l'ouvrage [5] aidera ceux qui étudient l'analyse des systèmes à découvrir de nombreuses particularités du travail du mathématicien dans les domaines appliqués.

Mais pour être véritablement utile aux chercheurs cet ouvrage doit tout de même indiquer des précédents. Aussi le dernier chapitre est-il consacré à quelques exemples concrets d'application de l'analyse des systèmes. Pour champ d'application des idées et méthodes de l'analyse des systèmes nous avons jeté notre dévolu sur le problème d'automatisation des projets. Ce domaine est certainement celui où les ordinateurs sont appelés à jouer un rôle particulier et où l'analyse des systèmes semble être l'instrument de prédilection.

Le sujet débattu est ample et les termes « automatisation du projet » doivent le restreindre quelque peu. Et pourtant l'« automatisation du projet » englobe aujourd'hui d'innombrables notions de nature variée, dont des notions purement technologiques: structure des banques de données, moyens d'automatisation des travaux graphiques, programmes de commande et même des langages de programmation. Mais nous glisserons là-dessus.

Le problème majeur étudié est celui de l'automatisation de l'étape initiale du projet, dite avant-projet. Notre attention sera essentiellement concentrée sur le choix des diverses options de l'ébauche du projet (ou comme on dit encore sur les contours du projet) ou l'élaboration d'un schéma général d'une unité économique. Ce problème est manifestement non seulement le plus ardu mais aussi la clef de voûte du projet. En effet, une erreur commise au départ ne peut être rectifiée ni par un perfectionnement de la technique graphique, ni par les méthodes de traitement des résultats de l'expérience, ni par les machines-outils à commande numérique, ni enfin par le perfectionnement des méthodes de calcul d'ingénieur. Elle ne peut non plus être compensée par la qualité des ordinateurs utilisés. L'avant-projet est en tout état de cause ce domaine où une bonne base scientifique, c'est-à-dire l'application des dernières méthodes de l'analyse, est susceptible de pallier à la relative carence des ordinateurs. Certes tout ce qui va être évoqué plus bas est impossible à réaliser sans ordinateur. Mais ce n'est pas à ce dernier de jouer le premier violon, c'est au perfectionnement des méthodes de l'analyse des systèmes.

Si nous apprenons aujourd'hui à aider le constructeur, le projeteur ou le planificateur à choisir correctement le prototype d'une future construction ou du projet d'une unité économique, alors la théorie de la décision apportera *ipso facto* une contribution décisive à l'automatisation du projet qui est si indispensable en cette période de brusque complication des constructions économiques et techniques, des technologies, etc. Nous sommes convaincus que l'utilisation d'un arsenal moderne de moyens de traitement de l'information et de méthodes scientifiques exigera une restructuration radicale des processus d'élaboration des projets et de planification ainsi qu'un perfectionnement des procédures décisionnelles.

Circonscrire tous les problèmes reliés aux processus décisionnels lors de l'élaboration de projets et de technologies est une tâche vaine quelle que soit la dimension de l'ouvrage. Aussi l'auteur s'est-il fixé un objectif moins ambitieux : composer un ouvrage appelé à montrer au lecteur que la mise au point d'un système de procédures de programmation doit reposer sur des principes généraux.

Il est une difficulté que l'auteur a constamment trouvé sur son chemin. Autant les chercheurs qui possèdent un appareil mathématique *ad hoc* se penchent rarement sur l'analyse de systèmes concrets, autant les analystes de systèmes concrets sont rarement armés d'un arsenal mathématique assez riche. Aussi l'auteur se trouve-t-il en butte au problème épineux d'écrire un ouvrage qui soit assez simple pour l'analyste de systèmes concrets et non rébarbatif pour les personnes qui élaborent les méthodes mathématiques de l'analyse des systèmes.

Cet ouvrage a son origine dans les cours donnés par l'auteur à des dates différentes à l'Université d'Etat de Moscou et à l'Institut de physique technique de Moscou. Les cours lus par l'auteur aux élèves ingénieurs de la faculté de calcul numérique et de cybernétique de l'Université de Moscou ont fortement influencé la conception de cet ouvrage. C'étaient des spécialistes qui avaient affaire à l'aspect « pratique » de l'analyse des systèmes. Les deux semestres de travail avec ce groupe ont convaincu l'auteur de la nécessité d'un cours synthétique intitulé *Introduction à l'analyse des systèmes*. Ce cours était appelé à rattacher de nombreuses disciplines enseignées généralement dans les universités: méthodes d'optimisation, éléments de recherche opérationnelle, théorie de la commande optimale, chapitres supplémentaires d'équations différentielles, aux procédures euristiques sans lesquelles est impossible l'analyse des systèmes plus ou moins complexes.

Le trait d'union est ici la simulation et le dialogue homme-machine, dont l'organisation prend progressivement la forme d'une branche scientifique autonome.

L'ouvrage est composé de quatre parties. La première, de trois chapitres, expose les bases méthodologiques de l'analyse des systèmes et montre comment les méthodes de la recherche opérationnelle et de la théorie de la commande ont inspiré une conception générale des systèmes cybernétiques et de l'analyse des systèmes.

La deuxième partie, de trois chapitres aussi, traite des méthodes de la théorie des perturbations.

La troisième partie, d'un chapitre, s'attache à montrer ce qu'on entend par « procédures euristiques ».

Quant à la quatrième et dernière partie qui comporte un seul chapitre, elle est consacrée, comme nous l'avons déjà signalé, à l'automatisation de l'élaboration des projets.

Cette articulation des sujets permet de cerner clairement la place

des mathématiques et du mathématicien dans l'analyse des systèmes et de dégager les éléments formels et non formels de cette discipline.

La structure de l'ouvrage, le choix et l'agencement des thèmes et notamment le caractère de sa présentation et de son interprétation sont le fruit de longues années de travail en commun et d'âpres discussions avec mes proches collègues du Centre de calcul de l'Académie des sciences d'U.R.S.S.: P. Krasnochtchekov, Yu. Pavlovski, A. Pétrov et autres. Yu. Evtouchenko, F. Erechko, A. Kononenko et Mme E. Stoliarova ont fourni une aide précieuse à l'auteur en relisant entièrement le manuscrit et en apportant au texte de nombreux perfectionnements. A toutes ces personnes l'auteur exprime sa sincère gratitude.

*N. Moïsséev*

Moscou, novembre 1983

CHAPITRE PREMIER

# MÉTHODES DE LA RECHERCHE OPÉRATIONNELLE ET ANALYSE DES SYSTÈMES

## § 1. Remarques introductives

La recherche opérationnelle est, ainsi qu'on l'a signalé dans l'avant-propos, l'une des principales sources de l'analyse des systèmes. Bien plus, les concepts et principes fondamentaux de l'analyse des systèmes approfondissent les idées de la théorie de la recherche opérationnelle et ses méthodes constituent aujourd'hui l'une des plus importantes branches de l'analyse des systèmes. Voilà pourquoi le premier chapitre de cet ouvrage est consacré à l'exposition des idées fondamentales de la recherche opérationnelle et à une interprétation de ces dernières qui soit nécessaire pour la suite.

Le terme « recherche opérationnelle » est apparu dans les années d'après-guerre lorsqu'on s'est aperçu à l'évidence que de nombreuses sphères de l'activité humaine avaient, malgré leurs distinctions qualitatives, un dénominateur commun : elles aboutissaient toutes au choix d'une méthode d'action, d'une variante de plan, de paramètres de construction, bref à la prise d'une décision et ce point commun a suffi à susciter l'élaboration d'une théorie cohérente et d'un système cohérent de méthodes. C'est dans ce contexte qu'a été forgé le terme « opération », un terme à vrai dire très général qui signifie toute action dirigée. En parlant de l'opération on lui associera toujours un responsable qui en conçoit le but et qui en est le bénéficiaire, et un analyste qui œuvre dans les intérêts du responsable et dont la tâche est de trouver un moyen d'utilisation des ressources (c'est-à-dire des possibilités du responsable) afin de réaliser l'objectif désigné. L'objectif de l'opération qui est généralement un élément exogène est supposé donné.

Dans cette position générale, la nouvelle discipline répondait aux besoins de nombreux domaines de l'activité humaine. A partir des années quarante, les problèmes de la recherche opérationelle font l'objet d'un nombre sans cesse croissant de travaux : travaux de mathématiques pures, travaux méthodologiques, travaux d'analyse d'opérations concrètes en économie, art militaire, agriculture, planification, etc.

La recherche opérationnelle fut promue au rang de discipline scientifique dans les années d'après-guerre mais ses bases furent jetées bien avant. Les travaux des mathématiciens soviétiques contribu-

èrent fortement à la formation des principes et des méthodes de la recherche opérationnelle.

Les problèmes liminaux de la recherche opérationnelle appartenaient aux branches les plus variées de l'activité humaine. Ainsi, D. Ventsel et V. Pougatchev développèrent la « théorie de l'efficacité des systèmes techniques » qu'ils armèrent de nombreux méthodes et principes de choix des paramètres qui réaliseraient le mieux les objectifs fixés par les constructeurs de ces systèmes. Des idées de même nature commencèrent à influencer sérieusement l'évolution de la théorie de la commande des systèmes techniques, discipline qui à la veille de la guerre entama l'étude des systèmes complexes de commande du genre pilote automatique. La détermination des caractéristiques constructives qui contribueraient à la meilleure réalisation de l'objectif fixé, par exemple au maintien du cap d'un avion, devient progressivement l'un des problèmes clefs de la théorie de la commande.

Des problèmes analogues se posaient en économie. A la fin des années 30 fut résolu le fameux problème de la coupe optimale, fut formulé le problème de transport, etc.

Dès 1927, F. Ramsey énonça le problème de répartition optimale des investissements. Signalons que les problèmes de la meilleure répartition des ressources furent toujours au centre des préoccupations en économie mais, de par leur complexité, les problèmes posés ne se prêtaient pas encore à une analyse efficace. Pour déployer un vaste front de recherches économico-mathématiques, on attendait la création et l'implantation de puissants ordinateurs.

A. Khintchine et B. Gnédenko commencèrent vers le milieu des années 30 à étudier une classe de problèmes probabilistes appelés par la suite problèmes des files d'attente. Cette orientation reçut une forte impulsion pendant la guerre vu la nécessité de planifier les opérations militaires et de dépenser rationnellement les ressources qui, en principe, sont très réduites.

Le nom même de recherche opérationnelle apparut en 1940 en Grande-Bretagne, où des équipes de scientifiques étudiaient les opérations militaires, puis atteignit rapidement les Etats-Unis.

La création des calculatrices au lendemain de la guerre enrichit considérablement l'arsenal numérique des mathématiques. Ceci eut des incidences immédiates sur le développement de toutes les théories liées à des problèmes pratiques concrets, donc sur la nécessité d'effectuer des calculs complexes et variés.

L'avènement des calculatrices fut un important facteur d'agrégation des divers problèmes décisionnels en une seule discipline scientifique appelée recherche opérationnelle.

Yu. Hermeyer joua un grand rôle dans la promotion de cette discipline en U.R.S.S. Son nom reste attaché à une meilleure compréhension de l'esprit de cette discipline, de la place de cette dernière

au sein de la science d'après-guerre (cf. [3]), à l'essor de méthodes mathématiques spéciales. Il introduisit également la notion nouvelle de « théorie de la recherche opérationnelle » pour souligner l'existence d'un principe conceptuel, c'est-à-dire d'une méthode générale en analyse des problèmes décisionnels de nature physique foncièrement différente.

Cette mise au point joua son rôle et fut indispensable dans la mesure où dans la littérature anglo-saxonne prévalait une approche pragmatique teintée d'ecclectisme: la recherche opérationnelle se présentait comme un amalgame de problèmes plus ou moins semblables justiciables des mêmes méthodes de résolution. Ce n'est qu'après les travaux de Hermeyer que l'on put parler de la recherche opérationnelle comme d'une discipline unie étudiant une classe de modèles de l'activité humaine. Dans cet ouvrage on utilise indifféremment « recherche opérationnelle » et « théorie de la recherche opérationnelle ». Toutes les fois qu'on emploira le terme « recherche opérationnelle » on sous-entendra le sens plus profond dont on vient de parler.

La nouvelle discipline scientifique ne put être traitée comme une discipline purement mathématique même si elle faisait une large part aux méthodes mathématiques et ouvrait de nombreuses voies aux mathématiques appliquées. Elle se penchait essentiellement sur des problèmes décisionnels complexes dans la résolution desquels les méthodes non formelles, le bon sens et les moyens de description — c'est-à-dire la formalisation mathématique des problèmes — tenaient une place aussi importante que l'appareil mathématique formel.

REMARQUE. L'intuition de l'analyste est capitale au niveau déjà de l'énoncé du problème qui est basé avant tout sur une analyse complète des données. Il doit non seulement comprendre le problème, mais le formuler dans des termes qui se prêtent à une analyse mathématique.

Ainsi, la recherche opérationnelle est une discipline synthétique dans laquelle se dégagent trois grandes directions dont l'une seulement ne rompt pas avec les applications traditionnelles des mathématiques. Ces orientations correspondent aux trois étapes suivantes qui se retrouvent dans toutes les recherches.

a) *Construction du modèle*, c'est-à-dire *formalisation du processus ou du phénomène étudié*. Cette formalisation consiste à décrire le processus en termes mathématiques. A ce niveau il est question de la construction d'un modèle du processus mais pas d'une opération. Un modèle peut servir à étudier plusieurs opérations.

b) *Description de l'opération — position du problème*. Le responsable définit le but de l'opération. Ce but est toujours supposé exogène à l'opération et doit encore faire l'objet d'une formalisation. L'analyste a pour tâche de procéder à une analyse des indéterminations et des contraintes et de formuler, de pair avec le responsable un *problème d'optimisation*

$$f(x) \Rightarrow \max, \qquad x \in G, \tag{1.1}$$

où $x$ est un élément d'un espace normé $E$ défini par la nature du modèle, $G \subset E$ un ensemble éventuellement très compliqué défini par la structure du modèle et par les particularités de l'opération étudiée. A cette étape donc, la recherche opérationnelle est traitée comme un problème d'optimisation. En fait la tâche de l'analyste est plus étendue. Après avoir analysé les contraintes imposées à l'opération, c'est-à-dire l'objectif fixé par le responsable et les inévitables indéterminations, l'analyste doit formuler l'objectif de l'opération en termes mathématiques. Le langage d'optimisation est naturel et commode mais pas unique. Au chapitre II nous verrons des opérations dans lesquelles le choix de la solution est lié à d'autres problèmes mathématiques. Il sera notamment question du choix des paramètres d'un pilote automatique qui permet à l'avion d'atteindre un point donné. Nous constaterons de plus que le choix des paramètres est effectué de manière à assurer la stabilité du vol de l'avion. Donc, la représentation de l'objectif sous la forme (1.1) n'est pas le seul procédé de formalisation. Mais il a l'avantage d'être commode dans la mesure où les méthodes d'optimisation sont suffisamment bien élaborées et le langage d'optimisation est, on le verra, assez universel.

c) *Résolution du problème d'optimisation.* A strictement parler, seule la troisième étape se rapporte aux mathématiques, bien que les deux premières soient impossibles à réaliser sans la participation d'un mathématicien (avec ses connaissances du langage mathématique et des possibilités de son appareil). De fines méthodes mathématiques peuvent être mises en jeu. La complexité du problème (due par exemple à la dimension du vecteur $x$ ou à la structure de l'ensemble $G$) ne permet pas souvent de se limiter à une étude purement mathématique de (1.1) et pour mener l'opération à son terme on doit faire appel à divers procédés euristiques. Signalons en passant que l'analyse non formelle pose des difficultés qui sont parfois déterminantes. Finalement les conditions posées et le caractère de la description du processus peuvent être des facteurs décisifs d'efficacité de l'analyse.

Une remarque s'impose à ce propos. L'un des plus grands mathématiciens russes A. Liapounov estimait qu'il était nécessaire de traiter tout problème physique une fois posé comme un problème de « mathématiques pures », c'est-à-dire de n'utiliser aucun raisonnement à caractère non formel. Ce point de vue est très difficile à appliquer en recherche opérationnelle. En effet, cette discipline fait constamment intervenir des raisonnements non formels. C'est pourquoi la vérification de la qualité d'une solution, sa conformité à l'objectif fixé constituent un très important problème de théorie.

La recherche opérationnelle a forgé sa propre terminologie et ses principes d'analyse. Etant donné que par opération on comprendra toute action orienté, par « modèle d'opération » on entendra un ensemble comprenant le responsable qui fixe l'objectif de l'opération, les ressources actives pour la réalisation de l'opération, l'ensemble

des stratégies, c'est-à-dire des moyens d'utilisation de ces ressources, et les critères ou procédés de comparaison des diverses stratégies visant à réaliser l'objectif de l'opération. Le critère ou, plus exactement, la maximalisation ou la minimisation de ses valeurs sont souvent adoptées pour objectif de l'opération.

Cette définition suscite parfois de notoires difficultés. En effet, le choix d'une stratégie est généralement soumis à de nombreuses contraintes et l'on a souvent intérêt à prendre l'une d'elles (par exemple, la satisfaction des besoins du client) pour objectif de l'opération. L'objectif de l'opération peut être réalisé de plusieurs manières et le critère sert à sélectionner la plus économique des stratégies admissibles, c'est-à-dire des stratégies vérifiant toutes les contraintes et réalisant le but de la commande.

Dans le même ordre d'idées, il est commode de définir spécialement la notion de modèle mathématique de l'opération ou ensemble de toutes les contraintes et conditions. Dans ce cas le critère n'est pas inclus dans le modèle. Cela signifie qu'une même stratégie, une même réalisation de l'opération peuvent être différemment appréciées. Cette terminologie est inspirée par la théorie de la commande. Elle est commode mais non universelle. Nous identifierons parfois les notions de critère et de but de la commande: nous ne serons pas très pointilleux sur les questions de terminologie. Le choix des termes sera dicté par la commodité de l'usage.

S'agissant des contraintes, on les classera en deux groupes: les contraintes physiques et les contraintes critérielles. Les secondes portent sur la construction ou le projet. Quand on projette un avion on veut, outre une économie maximale, que la vitesse de croisière ne soit pas par exemple inférieure à 800 km/h; quand on cultive une terre on cherche à en tirer une récolte maximale pour une nomenclature donnée. Ces contraintes ne sont pas très rigides. Elles sont du ressort du responsable et en principe elles peuvent être violées ou modifiées. En y renonçant on ne contredit pas les lois physiques du processus. Tel n'est pas le cas des contraintes physiques qui résultent des lois de conservation. Soit par exemple $q_i$ la norme d'arrosage, c'est-à-dire la quantité d'eau que nous devons fournir à une unité $x_i$ de surface irriguée. Alors

$$\sum_i q_i x_i \leqslant Q, \tag{1.2}$$

où $Q$ est la capacité du barrage.

Par ailleurs, l'aire totale $X$ des terres cultivables doit aussi être fixe, c'est-à-dire que les quantités $x_i$ doivent vérifier encore une contrainte:

$$\sum x_i \leqslant X. \tag{1.3}$$

Les conditions (1.2) et (1.3) ne peuvent en aucun cas être violées, car elles expriment des lois de conservation. Cette circonstance est susceptible de créer des difficultés de principe.

Signalons que les conditions (1.2) et (1.3) sont foncièrement différentes. Les quantités $x_i$ et $X$ de la condition (1.3) sont bien déterminées, quant aux quantités $q_i$ et $Q$ de la condition (1.2), elles sont aléatoires. Nous reviendrons sur cette question dans un prochain paragraphe de ce chapitre.

L'analyste est un important élément. En principe, il fait cause commune avec le responsable sans s'identifier à lui. Il possède une autre information sur la situation de l'opération. L'opération doit être étudiée du point de vue de l'analyste, à partir de sa propre information, information qui peut être actualisée par le responsable.

Nous concrétiserons les principes généraux et la terminologie au fur et à mesure que nous progresserons dans l'étude des problèmes de recherche opérationnelle.

## § 2. Quelques problèmes types de la recherche opérationnelle

Indiquons quelques représentants des classes de problèmes rencontrés par l'analyste.

a) *Le problème de transport*. Supposons qu'en des endroits $a_1$, $a_2$, ..., $a_n$ sont localisés des dépôts contenant des biens en quantités $X_1$, $X_2$, ..., $X_n$ respectivement. En $b_1$, $b_2$, ..., $b_m$ se trouvent des consommateurs auxquels il faut livrer ces biens en quantités $\geqslant Y_1$, $Y_2$, ..., $Y_m$ respectivement. Désignons par $d_{ij}$ le coût du transport d'une unité de marchandise entre $a_i$ et $b_j$.

Etudions l'opération de livraison de biens en quantités suffisantes pour satisfaire les besoins des consommateurs. Désignons par $x_{ij}$ la quantité de produits livrés de $a_i$ en $b_j$. Pour satisfaire les besoins des clients, il faut que $x_{ij}$ vérifient l'inégalité

$$\sum_i x_{ij} \geqslant Y_j. \tag{2.1}$$

Mais dans un dépôt on ne peut prendre plus de biens qu'il n'en contient. Donc les inconnues doivent encore vérifier le système d'inégalités:

$$\sum_j x_{ij} \leqslant X_i. \tag{2.2}$$

Il existe une infinité de moyens de satisfaire les conditions (2.1) et (2.2), c'est-à-dire de dresser un plan de livraison satisfaisant les besoins des consommateurs. Pour que l'analyste puisse opter pour une solution, c'est-à-dire indiquer des quantités $x_{ij}$, il faut formuler une certaine loi de sélection définie par un critère reflétant notre idée

subjective du but. Ceci nous donnera une des estimations possibles de la solution retenue.

On a déjà signalé que le critère n'était pas de la compétence de l'analyste, mais de celle du responsable. Dans ce problème un éventuel critère est le coût du transport, qui se définit visiblement par

$$J(x) = \sum_i \sum_j d_{ij} x_{ij}. \tag{2.3}$$

Le problème de transport se formule maintenant comme suit: on demande les quantités $x_{ij} \geqslant 0$ qui vérifient les contraintes (2.1), (2.2) et qui minimisent la fonction (2.3).

La contrainte (2.2) est la condition d'équilibre ou la loi de conservation, c'est-à-dire une condition d'ordre physique; la condition (2.1) est de toute évidence l'objectif de l'opération, puisque celui-ci consiste à satisfaire les besoins des consommateurs. Ces deux conditions constituent en fait le modèle de l'opération. La réalisation de l'opération dépendra du critère, c'est-à-dire du procédé choisi pour atteindre l'objectif de l'opération.

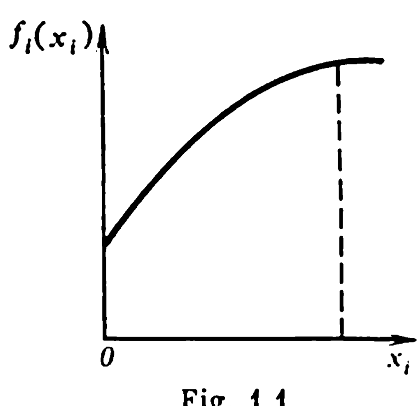


Fig. 1.1

Une telle différenciation s'explique par le fait qu'un même modèle d'opération (c'est-à-dire un modèle d'actions poursuivant le même but) peut donner lieu à des critères différents, c'est-à-dire à des procédés différents d'estimation de la voie à suivre pour réaliser l'objectif. Le critère peut donc prendre des aspects différents. Il peut se présenter soit comme un moyen de formalisation du but, soit comme un principe de choix d'actions admissibles, c'est-à-dire vérifiant les contraintes.

b) *Problème de répartition des engrais.* Soit à répartir une quantité donnée d'entrais entre $n$ cultures différentes. Supposons que le rendement $f_i\,(x_i)$ de la culture d'indice $i$ est une fonction de $x_i$ concave non linéaire, $x_i$ étant les quantités d'engrais utilisées par unité de surface (fig. 1.1). La récolte de la culture $i$ sera alors égale à $s_i f_i\,(x_i)$, où $s_i$ est l'aire occupée par la culture $i$. On admettra que l'aire totale est fixe, soit

$$\sum_{i=1}^{n} s_i \leqslant S, \tag{2.4}$$

où $S$ est un nombre donné. On admettra aussi que l'assortiment des cultures est bien défini, c'est-à-dire que

$$\frac{s_i f_i\,(x_i)}{s_1 f_1\,(x_1)} = \lambda_i, \quad i = 2, 3, \ldots, n, \tag{2.5}$$

où $\lambda_i$ sont des nombres donnés.

2*

Introduisons la contrainte

$$\sum_{i=1}^{n} s_i x_i \leqslant X, \tag{2.6}$$

où $X$ est la quantité disponible d'engrais.

En modifiant les quantités $x_i$ et $s_i$ sans violer les conditions (2.4), (2.5), (2.6), on obtiendra des variantes différentes de plan d'utilisation de l'aire $S$. Nous devons être en mesure de comparer ces plans. Considérons le critère suivant: désignons par $p_i$ le coût d'une unité de produit $i$, par $q$, le coût d'une unité d'engrais. Le bénéfice réalisé est égal au prix de vente de la récolte moins le prix d'achat des engrais, soit

$$\begin{gathered} J(x, s) = \sum_{i=1}^{n} (p_i s_i f_i(x_i) - q s_i x_i), \\ x = (x_1, \ldots, x_n), \quad s = (s_1, \ldots, s_n). \end{gathered} \tag{2.7}$$

Nous pouvons chercher un procédé de répartition des terres qui maximise la fonctionnelle (2.7) sous les contraintes (2.4), (2.5), (2.6). Signalons que la quantité $X$ peut également être traitée comme une inconnue.

c) *Problème d'irrigation et d'ensilage.* Considérons maintenant un problème plus complexe faisant intervenir des quantités aléatoires. Ce problème est une variante simplifiée du problème de répartition des investissements pour l'irrigation et la construction de silos. C'est un problème de prise de décisions en plusieurs étapes (en ce sens que l'on étudie un processus dynamique qui se déroule dans le temps), puisque la planification des investissements est effectuée pour plusieurs années à l'avance. Les facteurs aléatoires sont les conditions météorologiques qui confèrent à la récolte son caractère aléatoire. Désignons ces quantités aléatoires par $p$ et $q$ : $p$ étant le rendement des terres non irriguées, $q$, celle des terres irriguées ($q \geqslant p$ sous les mêmes conditions météorologiques). On admettra que les fonctions de répartition $F_p$ et $F_q$ de $p$ et $q$ sont connues. Désignons par $S(n)$ et $s(n)$ respectivement les aires des terres non irriguées et des terres irriguées pendant l'année $n$. On supposera que l'aire totale

$$S^*(n) = S(n) + s(n) \tag{2.8}$$

est connue. On admettra enfin que les besoins annuels $\Phi(n)$ en blé sont donnés.

La récolte de l'année $n$ sera visiblement une quantité aléatoire $\Phi^+(n) = pS(n) + qs(n)$, dont on peut calculer la fonction de répartition $F_{\Phi^+(n)}$.

La différence $\Phi^+(n) - \Phi(n)$ peut être soit $> 0$ soit $< 0$. Dans le premier cas, l'excédent de récolte est ensilé; dans le second, le défaut peut être puisé dans le silo. Cette quantité doit satisfaire des

relations évidentes. Ainsi, on ne peut ensiler plus de blé que les silos ne peuvent en contenir. Par ailleurs, la capacité des silos de réserve dépend du passé (de la préhistoire), c'est-à-dire de la quantité de blé puisée ou envoyée les années précédentes dans les silos. Le volume des silos dépend, quant à lui, des investissements alloués à leur construction. De la même façon, la quantité de blé retirée du silo dépend de la quantité disponible pendant l'année $n$, c'est-à-dire dépend de la préhistoire du processus. Désignons par $Q(n)$ la quantité de blé ensilée ou retirée du silo (dans le premier cas $Q(n) > 0$, dans le second, $Q(n) < 0$). Soient $R(n-1)$ la quantité de blé se trouvant dans le silo pendant l'année $n-1$ et $G(n)$ la capacité totale des silos pendant l'année $n$. Alors

$$Q(n) = \begin{cases} \min(\Phi^+(n) - \Phi(n), G(n) - R(n-1)), & \text{si } \Phi^+(n) \geqslant \Phi(n), \\ \max(\Phi^+(n) - \Phi(n), -R(n-1)), & \text{si } \Phi^+(n) < \Phi(n). \end{cases} \quad (2.9)$$

Les quantités $Q$, $R$, $G$ et $\Phi$ (qui sont exprimées dans les mêmes unités — en mètres cubes ou en tonnes) doivent manifestement toutes vérifier les relations dynamiques

$$G(n) = G(n-1) + \frac{x(n-1)}{C_x}, \quad (2.10)$$

$$R(n) = R(n-1) + Q(n), \quad (2.11)$$

où $x(n-1)$ désigne les capitaux investis dans la construction des silos, $C_x$, le prix de revient d'une unité de capacité des silos.

La quantité $\Phi^+(n)$ dépend de l'étendue des terres irriguées $s(n)$, laquelle est définie par la condition dynamique

$$s(n) = s(n-1) + \frac{y(n-1)}{C_y}, \quad (2.12)$$

où $C_y$ sont les dépenses par unité de terre irriguée, $y(n-1)$, les dépenses pour l'irrigation au cours de l'année $n-1$.

Les équations (2.8), (2.10), (2.11) et (2.12), où $Q(n)$ est définie par (2.9), décrivent le modèle mathématique du processus étudié. Les quantités $x(n)$ et $y(n)$ sont reliées par la contrainte

$$x(n) + y(n) = z(n), \quad (2.13)$$

où $z(n)$, qui est donnée, est la somme à investir dans la construction des silos et des systèmes d'irrigation. En se donnant d'une manière ou d'une autre les quantités $x(n)$ et $y(n)$ et l'état initial du système, c'est-à-dire les quantités $G(0)$, $R(0)$ et $s(0)$, on peut calculer la répartition du défaut $\Delta(n)$, soit

$$\Delta(n) = S(n)p + s(n)q - \Phi(n) - Q(n) \quad (2.14)$$

pour toute année $n$.

Passons maintenant à la discussion des éventuels critères d'efficacité (fonctions objectifs). Il est évident que plus l'espérance mathé-

matique de la valeur absolue du défaut sera petite et plus le système sera bon. Donc, pour critère d'évaluation du fonctionnement du système pour une année $n$, on peut prendre la valeur de cette espérance mathématique:

$$J(x, y) = E(|\Delta(n)|) = |\overline{\Delta(n)}|\ ^{*)}.$$

Mais comme le système est prévu pour plusieurs années, en désignant par $N$ l'horizon (le terme) de la planification, on peut retenir pour critère d'évaluation du système dans son ensemble

$$J_1 = \max_{1 \leqslant n \leqslant N} |\overline{\Delta(n)}|. \tag{2.15}$$

On peut remplacer le critère (2.15) par le critère

$$J_1^* = \overline{\max_{1 \leqslant n \leqslant N} |\Delta(n)|}, \tag{2.15'}$$

qui peut s'avérer plus commode au responsable. A noter qu'il majore toujours le critère (2.15):

$$J_1 \leqslant J_1^*,$$

mais $J_1$ se calcule en général bien plus aisément que $J_1^*$.

Le projet peut être évalué aussi par le critère

$$J_2 = \sum_{n=1}^{N} |\overline{\Delta(n)}|. \tag{2.16}$$

Les défauts strictement positifs ou strictement négatifs n'ont pas la même signification. Si $\Delta(n) > 0$, cela veut dire qu'une partie de la récolte est tout simplement perdue. Si $\Delta(n) < 0$, le blé ne suffit pas à couvrir les besoins et il faut alors l'importer.

Donc, pour nouveau critère on peut prendre la quantité

$$J_3 = \sum_{n} |\overline{\Delta(n)}|, \tag{2.17}$$

où la moyenne est prise sur les défauts strictement négatifs: $\Delta(n) <$ $< 0$.

On voit donc qu'une même opération peut être évaluée par des critères d'efficacité différents. Et comme les stratégies (ici la répartition des investissements) seront définies à partir de la condition

$$J_i \Rightarrow \min, \tag{2.18}$$

$i = 1, 2, 3$, alors à chaque $J_i$ sera associée sa propre stratégie, solution d'un problème (2.18); cette stratégie sera dite *stratégie optimale.*

d) *Problème de composition des horaires.* La théorie des horaires constitue un vaste domaine des mathématiques discrètes et de la thé-

---

*) Ici et dans la suite la barre désignera l'espérance mathématique de la quantité aléatoire correspondante.

orie de la recherche opérationelle. La composition d'un calendrier de travail, c'est-à-dire l'ordonnancement des tâches et l'allocation des ressources nécessaires à leur réalisation, occupe une place importante en planification et dans l'élaboration de projets d'unités techniques ou économiques complexes.

Illustrons ceci sur un problème fondamental de cette classe: trouver une répartition des ressources et un ordonnancement des tâches tels que le projet soit réalisé en un temps minimal. Supposons qu'un expert (un projeteur, un constructeur ou un planificateur) fixe les travaux $P_1, P_2, \ldots, P_n$ et les ressources nécessaires à la réalisation d'un projet. Ces ressources peuvent être de nature différente: des ouvriers (de telle ou telle qualification), des équipements, des matières premières, des fonds, etc. Donc, quand on dira que les ressources nécessaires sont données, on entend par là un certain *ratio* vectoriel, c'est-à-dire à chaque tâche est associé un vecteur énumérant les diverses ressources indispensables à la réalisation de cette dernière. Cependant la réalisation des travaux est souvent soumise à d'innombrables contraintes qui, en principe, se partagent en deux groupes.

Les *contraintes* ($\alpha$) décrivent l'interdépendance des travaux. Ces contraintes sont logiques. L'exemple le plus typique est fourni par les contraintes de type graphe: la réalisation du travail $P_i$ ne peut être entamée avant l'achèvement de certains travaux antérieurs. Un exemple suggestif nous est donné par la construction d'un immeuble: le toit ne peut être posé sans l'érection des murs, les murs, sans les fondations, etc.

Les contraintes ($\alpha$) peuvent être formulées dans le langage de la théorie des graphes. Convenons à cet effet de désigner les travaux par les sommets d'un graphe dont les arêtes orientées indiqueront l'ordre de réalisation des travaux. Les contraintes ($\alpha$) peuvent être de nature logique plus complexe *). La figure 1.2 représente un exemple de description des contraintes ($\alpha$).

Les *contraintes* ($\beta$) sont reliées au volume des ressources allouables à la réalisation du projet. Désignons par $v(t)$ le vecteur ressources investies dans le projet pendant l'année $t$**), par $u^i(t)$ la quantité de travail $P_i$ projeté pour l'année $t$ ($0 \leqslant u^i(t) \leqslant 1$), par $q^i(u^i(t))$ le vecteur ressources consacrées au travail $u^i(t)$. Ceci étant, les contraintes ($\beta$) peuvent être mises sous la forme

$$\sum_i q^i(u^i(t)) \leqslant v(t) \quad \forall t, \tag{2.19}$$

ou sous une forme analogue.

*) L'ordre des travaux peut par exemple être modifié, certains doivent être effectués simultanément à une certaine étape, etc.

**) Dans ce problème on traite le cas d'un temps discret: $t = 1, 2, 3, \ldots$ représente le numéro de l'intervalle temporel, par exemple le numéro de l'année, compté à partir du commencement du travail.

Les contraintes (β) sont les contraintes types rencontrées par l'analyste dans la résolution de presque tout problème de répartition.

Si donc les vecteurs $v(t)$ sont donnés, alors le plan de réalisation du projet se ramène au problème suivant: indiquer pour chaque intervalle de temps $t$ la liste et la quantité $u^i(t)$ de travaux à effectuer pour que le projet soit réalisé en un temps minimal. La quantité $u^i(t)$ est exprimée dans une échelle décimale, le temps $t$ est discret. Donc, le problème d'établissement du calendrier relève de la programmation discrète. Les solutions admissibles sont en nombre fini. Du

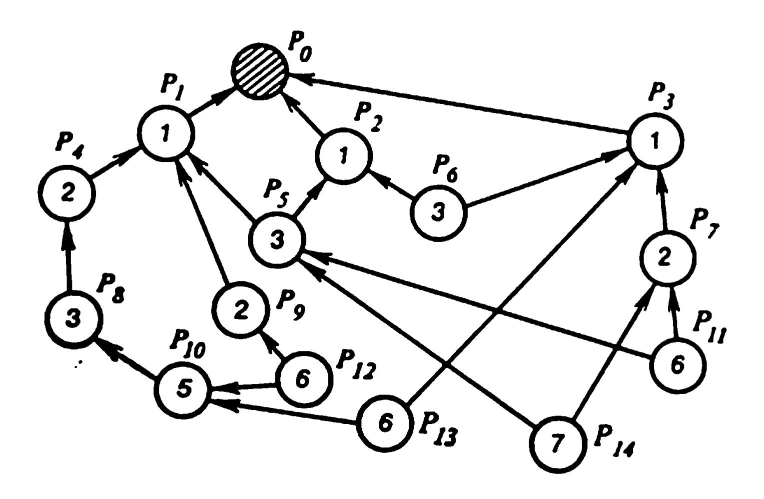



Fig. 1.2

point de vue des mathématiques « pures», ce problème ne pose aucune difficulté de principe: il peut être résolu par la méthode de tri complet. Mais si le nombre de travaux est assez élevé, ce procédé de résolution exige même sur un puissant ordinateur un temps astronomique et est pratiquement impossible. On démontre que la résolution par la méthode de tri complet des variantes possibles pour la composition d'un calendrier de 1000 travaux exige (sur un ordinateur IBM) une durée de temps égale à l'âge de notre Galaxie. Or de tels calendriers sont une affaire courante et la principale difficulté consiste à indiquer un ordonnancement des travaux ne violant pas les contraintes (α). Donc, dans les problèmes de la théorie des horaires, identiques à celui qui vient d'être décrit, on ne peut se passer de méthodes euristiques.

e) *Discussion.* Nous avons examiné quelques problèmes types rencontrés par l'analyste. Sur le plan mathématique, ce sont d'ordinaires problèmes de programmation. Le premier de ces problèmes, dit problème de transport, est l'un des plus élémentaires de la programmation linéaire. Le problème de répartition des engrais relève de la programmation non linéaire. Le troisième problème (c)), malgré son caractère probabiliste, appartient aussi à la programmation non

linéaire. Le problème d) enfin est un problème de programmation discrète.

Chacun de ces problèmes relève d'une branche des mathématiques et sa solution peut être acquise par des algorithmes divers bien élaborés. Nous ne développerons pas ces méthodes notoires qui font l'objet des cours d'optimisation des facultés de mathématiques appliquées et renvoyons le lecteur aux manuels respectifs (cf. [8]).

Le lecteur qui a suivi un cours d'optimisation peut penser que la recherche opérationnelle s'identifie aux branches des mathématiques impliquant la résolution de problèmes d'extrémum sous des contraintes spéciales. Mais une telle vision est erronée. C'est plutôt le contraire qui serait vrai. La théorie de la programmation mathématique, c'est-à-dire la théorie de la résolution des problèmes d'extrémum sous des contraintes, a vu le jour et s'est développée grâce essentiellement aux besoins de la recherche opérationnelle. C'est pourquoi de nombreux auteurs étrangers s'occupant des applications des mathématiques à la résolution de problèmes économiques ou techniques traitent les problèmes de programmation linéaire, non linéaire et discrète comme relevant non pas des branches utilisées en recherche opérationnelle, mais comme une partie intégrante de cette discipline. C'est un autre point de vue diamétralement opposé. A mon sens, toutes les branches de la programmation mathématique sont en réalité des branches mathématiques et leur développement est impossible si les problèmes de programmation mathématique ne sont pas considérés comme un objet de recherches mathématiques.

La programmation mathématique et les autres méthodes de résolution des problèmes d'extrémum constituent l'ossature de l'appareil de la recherche opérationnelle. Mais la théorie de la recherche opérationnelle ne peut en aucun cas être confinée à la résolution des problèmes d'extrémum. Bien plus, la recherche opérationnelle n'est pas une discipline de mathématiques pures et les principales difficultés de l'analyse d'opérations concrètes ne sont pas en principe d'ordre mathématique.

La discussion des problèmes exhibés dans ce paragraphe montre que le premier pas consiste à formaliser l'opération, à la décrire en termes mathématiques. De cette formalisation dépend tout le succès de la recherche. Une description simple simplifie l'analyse, mais si elle n'est pas assez conforme à la réalité, elle peut conduire à des résultats douteux. De même, une description trop fidèle à la réalité, prenant en considération les moindres détails du processus, peut provoquer une perte de temps machine qui n'est pas justifiée par la haute précision du résultat. En un mot, au niveau déjà de la composition du modèle, l'analyste qui, en principe, est un mathématicien, doit exercer son expérience, sa maîtrise, ses capacités à saisir le fond du problème et à s'imaginer clairement le but de la recherche. On voit donc que la première étape se démarque nettement des mathé-

matiques traditionnelles, néanmoins seul peut surmonter ces difficultés un homme qui a bien conscience des possibilités de son appareil, c'est-à-dire un mathématicien de facto et non de jure.

REMARQUE. Ces derniers temps on tente de dissocier les fonctions du programmeur et du « poseur » de problèmes. Cette opération doit être conduite avec la plus grande circonspection. Certes il est parfois nécessaire à une certaine étape de procéder à cette dissociation et la partie programme peut effectivement être confiée à un spécialiste en programmation, surtout s'il s'agit de problèmes d'organisation des systèmes de programmes, des programmes de gestion, etc. Mais pour assurer le succès de la recherche, il est absolument nécessaire que l'analyste soit à la fois un mathématicien et un spécialiste initié à toutes les subtilités de la matière étudiée. En d'autres termes, l'analyste doit être un mathématicien professionnel qui confine son activité dans l'étude d'un phénomène concret par les moyens des mathématiques.

Les choses se compliquent davantage au niveau de la définition du critère et de la comparaison des diverses stratégies. Il arrive fréquemment que l'opération s'évalue par des critères différents. On parle alors d'indétermination des buts. Cette indétermination est impossible à lever par des méthodes formelles et il faut alors recourir à des recherches et des hypothèses supplémentaires. Au paragraphe suivant on examinera ces hypothèses supplémentaires et certaines méthodes de représentation de l'information qui nous permettront dans bien des cas de surmonter cette indétermination.

La définition du critère donne lieu à d'autres difficultés liées au caractère stochastique des opérations. Supposons que le critère est de la forme

$$J(x, \xi) = f(x, \xi), \tag{2.20}$$

où $\xi$ est un paramètre aléatoire de loi de répartition connue, $x$, le vecteur des caractéristiques de construction du système projeté. Cela signifie que l'efficacité du système projeté dépend non seulement des paramètres de construction choisis par le projeteur, mais aussi de facteurs qui échappent à son contrôle. Le constructeur essaye de choisir les paramètres de manière à maximiser la valeur du critère. Or l'opération

$$J(x, \xi) \Rightarrow \max_x \tag{2.21}$$

n'a un sens que si l'on fixe le paramètre $\xi$. La résolution du problème (2.21) nous donne alors une fonction $\hat{x}(\xi)$, c'est-à-dire qu'à chaque valeur de $\xi$ est associée une valeur $x = \hat{x}(\xi)$. Donc, la stratégie $\hat{x}(\xi)$ est optimale dans le cas seulement où au moment de la prise de décision on connaît la valeur du paramètre aléatoire $\xi$. Dans le cas contraire on ne sait pas quel vecteur $x$ prendre.

Si la construction projetée est à usage répété, il est plus logique de prendre des valeurs des paramètres de construction qui maximisent

l'espérance mathématique du critère, c'est-à-dire qui sont solution du problème

$$I_1(x) = \overline{f(x, \xi)} \Rightarrow \max_x. \tag{2.22}$$

Cependant cette affirmation n'est pas absolue et en d'autres circonstances on utilisera d'autres méthodes. Par exemple, la variance du critère peut jouer un certain rôle. Peut-être aurons-nous parfois intérêt à être moins transigeants sur la valeur de l'espérance mathématique pour réduire la dispersion des résultats, c'est-à-dire la valeur de la variance

$$I_2(x) = \overline{(f(x, \xi) - \overline{f(x, \xi)})^2}. \tag{2.23}$$

A noter que l'espérance mathématique $\overline{f(x, \xi)}$ est une fonction de $x$:

$$\overline{f(x, \xi)} = f^*(x). \tag{2.24}$$

Son calcul peut s'avérer assez laborieux. Il faut d'abord se donner le vecteur $x$ et seulement ensuite prendre la moyenne, opération qui peut aussi exiger une importante perte de temps machine. Nous sommes maintenant en mesure de dire à quel point sera fastidieuse la procédure de recherche des valeurs extrémales de la fonction $f^*(x)$. Nous reviendrons dans un prochain chapitre sur la recherche des valeurs extrémales de fonctions du type (2.24).

Les difficultés soulevées par la résolution du problème (2.22) nous contraignent souvent à remplacer ce problème par un autre. A noter que la résolution du problème (2.21) pour une valeur fixe du paramètre aléatoire $\xi$ est parfois très simple, c'est pourquoi le problème de trouver

$$\max_x f^*(x) = f^*$$

est remplacé par celui de trouver

$$\overline{\max_x f(x, \xi)} = \hat{f}.$$

Comme $\hat{f} \geqslant f^*$ (ceci résulte de l'inégalité évidente $\max(f_1(x) + f_2(x)) \leqslant \max f_1(x) + \max f_2(x)$), on obtient une majoration utile. On peut se prévaloir d'autres considérations pour choisir la fonction objectif.

Si donc l'on envisage d'utiliser plusieurs fois la construction projetée, le problème du choix d'un critère de la forme (2.22) ou (2.23) diffère peu des situations ne faisant pas intervenir les paramètres aléatoires. Les difficultés supplémentaires qui apparaissent sont dues au calcul plus volumineux de la fonction objectif, puisque nous devons procéder à une moyennisation.

La situation est bien plus compliquée quand il est question du choix de paramètres de construction ou d'un plan appelés à être utilisés une seule fois. Nous avons eu affaire à un pareil cas en discutant la répartition des investissements pour la construction de silos et pour l'irrigation. L'information sur les caractéristiques statistiques n'a alors aucun sens formellement: quelle que soit la probabilité que $\xi = 10^{10}$ ou $\xi = 10^{-10}$ on ne peut rien dire de la valeur de la fonctionnelle et, à strictement parler, le choix de $x$ (et la valeur de la fonction $f(x, \xi)$) peut être arbitraire. Nous devons donc postuler une hypothèse, introduire une fonction de risque, évaluer les « chances », etc. C'est au fond ce que nous avons fait en déterminant la répartition « optimale » des investissements. L'hypothèse retenue était en gros la suivante: la fonction objectif est choisie comme si notre système était destiné à fonctionner plusieurs fois. Cette hypothèse n'est pas dénuée de sens, puisque ces investissements nous permettront de créer des biens qui nous serviront plusieurs années.

## § 3. Indétermination des objectifs

L'analyse des opérations envisagées dans le paragraphe précédent nous a conduits à des problèmes d'optimisation rigoureusement posés en ce sens qu'ils ne donnaient lieu à aucune indétermination. Les problèmes stochastiques, c'est-à-dire les problèmes faisant intervenir des quantités ou des fonctions aléatoires, ne seront pas rapportés aux problèmes contenant des facteurs indéterminés: si pour fonction objectif on peut prendre les valeurs moyennes de certaines quantités, alors la recherche d'une stratégie optimale se ramène, on l'a vu, à un ordinaire problème d'optimisation sans la moindre hypothèse supplémentaire.

Mais les problèmes sans indéterminations sont plutôt une exception qu'une règle: une description adéquate du problème contient pratiquement toujours des indéterminations de nature diverse traduisant la situation réelle de l'analyste, savoir que ses connaissances sont relatives et imprécises. En recherche opérationnelle, il est d'usage de distinguer trois sortes d'indéterminations: l'indétermination des buts, l'indétermination de nos connaissances sur le milieu ambiant (indétermination de la nature) et l'indétermination du comportement de l'adversaire ou du partenaire. Considérons à tour de rôle ces trois types d'indéterminations et essayons de comprendre ce dont l'analyste a nécessairement besoin pour analyser les problèmes respectifs par les instruments mathématiques.

Au paragraphe précédent, nous avons envisagé des situations dans lesquelles le choix des stratégies se ramenait à la détermination de valeurs extrémales de fonctions. Mais, au niveau déjà de la discussion de ces problèmes d'extrémum, nous avons attiré l'attention du

lecteur sur l'adéquation d'autres approches: pour critères on pouvait prendre d'autres fonctions objectifs.

Tout au début de ce chapitre, on a signalé qu'en recherche opérationnelle la fonction objectif était exogène. A l'issue de l'analyse de l'opération on choisit un moyen de réalisation du but, c'est-à-dire une stratégie. Mais la désignation de l'objectif, du critère, c'est-à-dire la formalisation de l'objectif (le choix de la fonction objectif), est toujours ou presque un problème épineux. Considérons de nouveau le problème de répartition des engrais. Nous avons utilisé un critère (cf. (2.7)) qui en réalité est une combinaison de deux critères: le prix de vente de la récolte et le prix d'achat des engrais. La tendance naturelle de l'analyste est de trouver une stratégie qui maximise le revenu (le prix de vente de la récolte) et minimise les dépenses (le prix d'achat des engrais). Si l'on se place dans l'optique de l'analyste, au lieu du problème (2.7), on aura à résoudre le problème

$$\begin{gathered} f(x) \Rightarrow \max, \\ -F(x) \Rightarrow \max \qquad (F(x) \Rightarrow \min), \end{gathered} \tag{3.1}$$

où $f(x)$ et $F(x)$ sont des fonctions caractérisant les revenus et les dépenses respectivement.

A la différence du problème envisagé dans le paragraphe précédent, ce problème n'admet pas en principe de solution. En effet, plus les dépenses $F(x)$ seront élevées, plus la récolte (donc $f(x)$) le sera.

Donc les deux objectifs sont contradictoires. A noter que cette situation se résume bien par l'expression: obtenir le maximum de produits avec le minimum de dépenses. Cette expression, même si elle est dénuée de toute signification scientifique — en effet, le minimum de dépenses est nul et avec des dépenses nulles on ne peut réaliser aucune tâche valable — elle n'en traduit pas moins correctement les intérêts du responsable. La situation envisagée est typique: elle montre que même s'il connaît les objectifs (les désirs) du responsable, l'analyste n'est pas encore en mesure de s'acquitter de sa « tâche principale », c'est-à-dire la résolution du problème d'optimisation. Les guillemets ont été mis spécialement pour souligner que la formalisation est une étape de révolution non moins importante et souvent la plus compliquée. De plus la « tâche principale » n'est pas nécessairement la résolution du problème d'optimisation, ce peut être le choix d'une solution satisfaisant telle ou telle condition.

Pour ramener un problème de recherche opérationnelle à un problème classique d'optimisation, il faut encore formuler les hypothèses supplémentaires ne résultant pas de la position du problème. Au paragraphe précédent, on a introduit comme hypothèse la fonction objectif

$$J = f(x) - F(x).$$

Mais comment formuler un seul objectif si les critères sont nombreux :

$$f_1(x) \Rightarrow \max, \quad f_2(x) \Rightarrow \max, \quad \ldots, \quad f_n(x) \Rightarrow \max,$$

et les moyens de les réaliser se trouvent entre les mains d'une seule partie ? Même si les mathématiques sont incapables de répondre de façon unique à cette question, elles nous fournissent les éléments nécessaires pour prendre une décision et faire un choix convenable. Tel est le problème de l'indétermination de l'objectif. Ce problème est inhérent à tout important projet technologique et économique. Par exemple, il est parfaitement logique qu'un constructeur aéronautique aspire à fabriquer un avion qui soit le plus rapide, le plus fiable, qui atteigne la plus grande altitude et qui soit aussi le moins coûteux. Or, il est en principe impossible de réaliser tout ceci simultanément. Une construction réelle est toujours une sorte de compromis, de combinaison des qualités requises. Mais le constructeur ne sait pas à priori lesquelles. Et c'est en cela que consiste le problème fondamental des multicritères (indéterminations des objectifs).

Ainsi l'indétermination des objectifs implique nécessairement l'introduction d'hypothèses supplémentaires si l'on veut formuler univoquement l'objectif de l'opération.

A noter que la recherche opérationnelle a commencé peu à peu à se transformer en une discipline synthétique s'appuyant non seulement sur un puissant appareil mathématique, mais aussi sur de nombreuses méthodes de levée des indéterminations. Le problème de prise de décision en présence d'indéterminations devient progressivement le problème central qui à proprement parler a fait de la recherche opérationnelle une théorie autonome. Cette théorie étudie en particulier les diverses méthodes de levée des indéterminations, les hypothèses nécessaires à cela et les propriétés des solutions les satisfaisant.

Attardons-nous ici sur certaines des méthodes les plus usitées pour lever les indéterminations et discutons le cas où l'analyste a à choisir une ligne d'action (le vecteur $x$) qui maximise simultanément les fonctions

$$f_1(x), \quad f_2(x), \quad \ldots, \quad f_n(x).$$

a) « *Convolution* » *linéaire*. Au lieu de $n$ critères $f_i$ on propose de considérer un seul critère de la forme

$$F(x) = \sum_{i=1}^{n} c_i f_i(x), \tag{3.2}$$

où $c_i$ sont des nombres $> 0$ normés par un procédé quelconque (par exemple $\sum_{i=1}^{n} c_i = 1$).

Ce procédé de « convolution » (de réduction) introduit en fait une relation d'équivalence des divers critères (des fonctions objectifs), puisque les coefficients $c_i$ indiquent de combien varie la fonction objectif $F(x)$ lorsque le critère $f_i(x)$ varie d'une unité: $c_i = \partial F/\partial f_i$.

Les coefficientes $c_i$ qui sont le résultat d'une expertise traduisent la valeur attachée par le responsable à la concession faite. Donc, le compromis consiste en une hiérarchisation des objectifs, qui, combinée au choix des coefficients de pondération, forme l'hypothèse supplémentaire qui permet de ramener le problème à plusieurs critères à un problème à un seul critère défini par la formule (3.2).

Cette hiérarchisation est visiblement loin d'être une méthode universelle pour lever l'indétermination des objectifs.

b) *Utilisation des indices de contrôle.* Dans les problèmes de planification et de projet on donne très souvent un système de *ratios* $f_1^*, f_2^*, \ldots, f_n^*$. Cela signifie par exemple que les paramètres de la future construction doivent maximiser des fonctions $f_i(x)$ telles que $f_i(x) \geqslant f_i^*$, $i = 1, 2, \ldots, n$.

Dans de tels cas il est commode de représenter la fonction objectif sous la forme

$$F(x) = \min_i \frac{f_i(x)}{f_i^*} \tag{3.3}$$

et de chercher le vecteur $x$ qui maximise $F(x)$. Le critère a été choisi sous la forme (3.3) pour une raison assez simple. Si le vecteur $x$ est donné, la quantité $F(x)$ représente la valeur du plus mauvais des indices $f_i(x)$. Donc, la condition $F(x) \Rightarrow \max$ désigne le choix d'un système de paramètres de construction $x$ qui maximise le rapport de la valeur réelle à la valeur de contrôle du $i$-ième critère. Si les valeurs de $f_i^*$ ne sont pas fixées, elles peuvent être déterminées par un devis.

Les critères de la forme (3.2) présentent l'important avantage suivant. Supposons que les contraintes imposées au choix des composantes du vecteur $x$ sont linéaires:

$$\sum a_s^j x^s \leqslant b_j, \tag{3.4}$$

ainsi que les fonctions $f_i(x) = \sum d_s^i x^s$. Il est alors évident que le problème du choix à l'aide du critère (3.2) se ramène à un problème de programmation linéaire: définir le maximum de la forme linéaire

$$F(x) = \sum_s \sum_i c_i d_s^i x^s$$

sous les contraintes linéaires (3.4).

Sous ces conditions, les critères (3.3) sont aussi linéaires. Ceci se démontre sans peine par l'introduction de la nouvelle variable

$$V = \min_i \frac{f_i(x)}{f_i^*}.$$

Aux contraintes (3.4) s'ajoutent de toute évidence les contraintes

$$f_i(x) \geqslant V f_i^*, \qquad i = 1, \ldots, n, \tag{3.5}$$

et l'on est conduit au problème de programmation linéaire suivant: déterminer par rapport à $x$ le maximum d'un scalaire $V$ vérifiant les contraintes (3.4) et (3.5).

c) *Méthode élémentaire de levée des indéterminations des objectifs.* Soit donné de nouveau un système d'indices de contrôle $f_i^*$ tels que

$$f_i(x) \geqslant f_i^*, \qquad i = 1, \ldots, n. \tag{3.6}$$

Supposons de plus que l'un des critères $f_i$, par exemple $f_1(x)$, est considéré comme essentiel. On est alors de nouveau conduit à un problème à un seul critère

$$f_1(x) \Rightarrow \max$$

sous les contraintes (3.6).

Cette réduction à des problèmes à un seul critère est probablement la plus simple et la plus couramment utilisée par les ingénieurs. La tâche du constructeur (du projeteur) consiste seulement à définir les bornes admissibles des indices utilisés.

d) *Introduction d'une métrique dans l'espace des fonctions objectifs.* Considérons encore une hypothèse souvent utilisée. Supposons que nous ayons résolu un système de problèmes à un seul critère

$$f_i(x) \Rightarrow \max, \qquad i = 1, 2, \ldots, n,$$

et que nous ayons trouvé dans le $i$-ième problème le vecteur $x = x_i$ maximisant le critère $f_i(x)$:

$$f_i(x_i) = \hat{f}_i, \qquad i = 1, \ldots, n. \tag{3.7}$$

L'ensemble des quantités scalaires $\hat{f}_i$ définit dans l'espace des critères un point que l'on appellera *point de maximum absolu.* Si les vecteurs $x_i$ sont distincts, il n'existe pas de choix permettant d'atteindre ce point: le point $(\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_n)$ est inaccessible dans l'espace des critères. Introduisons maintenant une matrice $R = (r_{ij})$ définie positive. La quantité scalaire

$$h = \sqrt{\sum_{i,j} (f_i(x) - \hat{f}_i)\, r_{ij}\, (f_j(x) - \hat{f}_j)} \tag{3.8}$$

définit alors dans l'espace des critères une distance du point correspondant au vecteur donné $x$ au point de maximum absolu. Dans le

cas particulier où $R$ est la matrice unité, la quantité

$$h = \sqrt{\sum_i (f_i(x) - \hat{f}_i)^2} \qquad (3.9)$$

représente la distance euclidienne du point $(f_1(x), f_2(x), \ldots, f_n(x))$ au point $(\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_n)$ dans l'espace des critères.

Pour nouveau critère scalaire on peut prendre la fonction (3.8). Sa minimisation fournit une information utile à l'analyste: elle indique nos possibilités limites d'atteindre le maximum absolu.

L'introduction de tels critères correspond aussi à des hypothèses dont l'adoption est laissée à l'appréciation de l'analyste. Ces critères ne sont pas meilleurs que ceux examinés plus haut bien qu'ils traduisent certaines propriétés du problème: ils sont porteurs d'une information utile au responsable.

e) *Compromis de Pareto.* Quand on a affaire à des problèmes à plusieurs critères, on essaye naturellement de trouver des méthodes pour les ramener à d'ordinaires problèmes à un seul critère, car pour ces derniers, et si de surcroît la fonction objectif est assez régulière, il existe des procédures de résolution assez bien élaborées. Ces méthodes doivent visiblement être non formelles, car elles ne peuvent être acquises par la résolution d'un quelconque problème mathématique. Nous avons déjà examiné plusieurs méthodes identiques basées sur la réduction des critères. La signification des méthodes de réduction des critères exposées est assez évidente: nous avons remplacé un problème par un autre et de plus l'adéquation de cette substitution constituait les nouvelles hypothèses.

Mais il existe d'autres méthodes d'analyse des problèmes à plusieurs critères, consistant à réduire l'ensemble des variantes initiales, c'est-à-dire à éliminer les solutions qui à priori sont mauvaises. Voyons l'une de ces méthodes proposée par l'économiste italien V. Pareto en 1904.

Supposons que nous ayons choisi une solution. Désignons-la par $x^*$ et admettons qu'il existe un autre vecteur $\hat{x}$ tel que pour tous les critères $f_i(x)$ l'on ait

$$f_i(\hat{x}) \geqslant f_i(x^*), \qquad i = 1, \ldots, n, \qquad (3.10)$$

l'une au moins de ces inégalités étant stricte.

Il est évident que la solution $\hat{x}$ est préférable à $x^*$. Donc, tous les vecteurs $x^*$ vérifiant (3.10) doivent être immédiatement exclus de l'analyse. Il vaut la peine de procéder à une comparaison et de ne soumettre à une analyse non formelle que les vecteurs $x^*$ pour lesquels il n'existe pas un $\hat{x}$ tel que les inégalités (3.10) soient réalisées pour tous les critères. L'ensemble de toutes ces valeurs de $x^*$ s'appelle *ensemble de Pareto* et le vecteur $x^*$, *vecteur inaméliorable des ré-*

*sultats* (ou *vecteur de Pareto*) si de $f_i(\hat{x}) \geqslant f_i(x^*)$, $\forall i$, il résulte que $f_i(\hat{x}) = f_i(x^*)$ (cf. [4]).

Supposons que les objectifs du responsable sont définis par deux fonctions:

$$f_1(x) \Rightarrow \max, \qquad f_2(x) \Rightarrow \max.$$

Alors à chaque valeur admissible de la variable $x$ est associé un point du plan $(f_1, f_2)$ (fig. 1.3), et les égalités

$$f_1 = f_1(x), \qquad f_2 = f_2(x)$$

représentent l'équation paramétrique d'une courbe *abcd* de ce plan. L'ensemble de Pareto ne contient pas toute la courbe. Ainsi la portion *bc* n'appartient visiblement pas à l'ensemble de Pareto, puisque

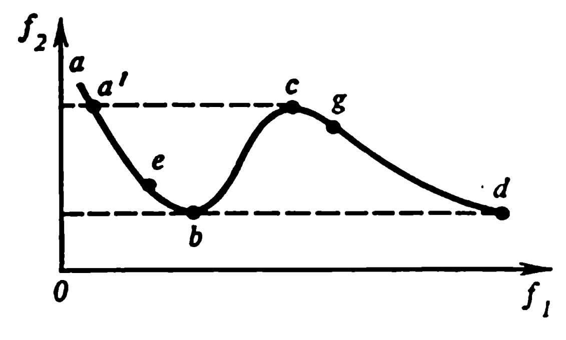



Fig. 1.3

$f_2$ croît avec $f_1$. Donc, sur cette portion l'accroissement de la variable $x$ se traduit par un accroissement simultané des deux fonctions objectifs et par suite de telles solutions doivent être immédiatement exclues de toute analyse ultérieure.

On exclura de même la portion $a'b$, car pour chacun de ses points $e$ il existe un point de la portion $cd$ en lequel les fonctions $f_1$ et $f_2$ prennent des valeurs supérieures à celles prises en $e$. Donc seules peuvent appartenir à l'ensemble de Pareto les portions $aa'$ et $cd$, le point $a'$ étant à exclure.

En théorie de la décision, le principe de Pareto consiste à ne choisir pour solution que le vecteur $x$ qui appartient à l'ensemble de Pareto. Le principe de Pareto ne distingue pas une solution unique, il ne fait que restreindre l'ensemble des alternatives. Le choix définitif revient au décideur. Mais en construisant l'ensemble de Pareto, l'analyste facilite la procédure de choix de la solution.

Le principe de Pareto joue un rôle très important dans l'automatisation de l'élaboration des projets. Considérons par exemple le projet d'un système d'alimentation en eau. La construction de ce système permettra de fournir de l'eau à de grosses unités industrielles et agricoles et d'augmenter de la sorte leur rendement. Mais ceci engendre une foule d'effets négatifs. La grande surface du barrage qui est

indispensable à la régulation du fonctionnement du système hydraulique provoque des effets de stagnation, une importante quantité d'eau est perdue par évaporation, etc. Par ailleurs, la réduction du débit du système fluvial dégrade les conditions de navigation et la pisciculture, quant à la construction de complexes industriels, elle accroît la pollution et par conséquent souille l'eau destinée à l'irrigation des champs, etc. En un mot, la situation dépend de plusieurs critères et les buts du projeteur peuvent être mis sous la forme

$$f_i(x) \Rightarrow \max, \qquad i = 1, \ldots, n.$$

Le projeteur est dans l'obligation de trouver un compromis. Et un moyen de le faire est de construire l'ensemble de Pareto dont l'étude nous fournira une riche information. Le décideur voit en particulier combien « coûte » l'accroissement d'un indice, comment il se répercute sur les autres dont les valeurs se détériorent irrémédiablement. Cet ensemble est en principe de nature très complexe et son analyse par des méthodes intuitives est peu probable.

Mais en plus de critères $f_i(x)$, le projeteur dispose assez souvent d'un critère général $F(x)$, qui est parfois formalisé et explicité. Ce peut être par exemple le coût du projet. Dans ce cas l'analyste a la possibilité de résoudre le problème jusqu'au « bout ». Il lui suffit à cet effet de définir le vecteur $x$ qui donne la solution du problème $F(x) \Rightarrow \max$ pour $x \in P_G(f_1, \ldots, f_n)$, où $P_G(f_1, \ldots, f_n)$ est l'ensemble de Pareto pour les fonctions $f_1, \ldots, f_n$ sur l'ensemble $G$ des vecteurs $x$ admissibles. Par exemple, dans le cas d'un système d'alimentation en eau, l'ensemble $G$ est défini par une répartition de l'eau entre les unités $x_i$, telle que la quantité d'eau soit inférieure au débit $Q(x)$.

L'introduction d'un critère « général » $F(x)$ et sa maximisation sur l'ensemble de Pareto est aussi une hypothèse, puisque nous avons particularisé un critère parmi l'ensemble $f_1, \ldots, f_n, F$.

REMARQUE. Les méthodes d'analyse séquentielle et d'élimination des variantes non concurrentielles, les méthodes de contraction successive de l'ensemble des alternatives jouent un grand rôle dans les problèmes décisionnels. Elles sont prépondérantes dans la formation des procédures de choix. Nous les aborderons en détail dans un prochain chapitre. Le principe de Pareto en fait précisément partie.

f) *Sur les méthodes numériques de construction de l'ensemble de Pareto.* La construction approchée de l'ensemble de Pareto est un problème très important et complexe de l'analyse numérique. L'importance des méthodes d'analyse de l'ensemble de Pareto croît sans cesse avec l'élargissement du champ des problèmes étudiés par l'analyse des systèmes (par exemple, avec l'apparition des problèmes d'automatisation des projets). Mais jusqu'à ces derniers temps ces problèmes ont fait l'objet de très peu d'attention et les méthodes numériques de construction de points de l'ensemble de Pareto ne sont qu'à

leurs balbutiements. Expliquons l'objet de ces problèmes sur quelques exemples simples.

Commençons par le cas élémentaire de deux critères. Soit donné le problème

$$f_1(x) \Rightarrow \max, \quad f_2(x) \Rightarrow \max, \quad x \in G_x. \tag{3.11}$$

A chaque point $x \in G_x$ les fonctions

$$f_1 = f_1(x), \qquad f_2 = f_2(x) \tag{3.12}$$

associent un point $f \in G_f$ du plan des critères (fig. 1.4). Les relations (3.12) définissent une application de l'ensemble $G_x$ sur $G_f$.

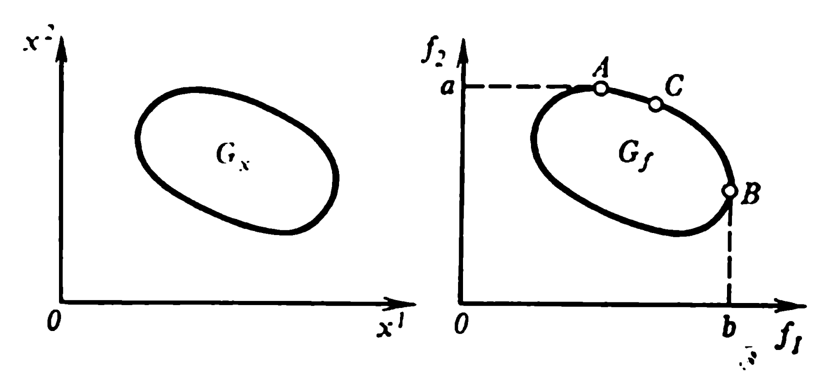



Fig. 1.4

L'ensemble $G_f$ s'appelle *ensemble permis* ou *ensemble des possibilités limites*. L'étude de la structure de cet ensemble peut être très utile dans les problèmes de projection et de planification. C'est pourquoi on a commencé ces dernières années à étudier systématiquement les possibilités de construction des ensembles permis. Signalons que l'ensemble de Pareto n'est qu'une partie de la frontière de l'ensemble permis. Sur la figure 1.4 l'arc *ACB* est un ensemble de Pareto.

La construction approchée de l'ensemble de Pareto se ramène à la résolution successive d'une série de problèmes de programmation mathématique. Décrivons un éventuel schéma de calcul. Fixons des valeurs désirées des critères $f_1$ et $f_2$ :

$$f_1 = C_1, \qquad f_2 = C_2.$$

Les valeurs $C_1$ et $C_2$ doivent appartenir à l'ensemble permis.

REMARQUE. Dans le cas général, la détermination d'un point intérieur de l'ensemble permis est un problème qui n'est pas toujours simple.

Considérons maintenant deux problèmes d'optimisation.

$$\text{I}: \; f_1(x) \Rightarrow \max, \quad x \in G_x, \; f_2(x) = C_2; \qquad \text{II}: \; f_2(x) \Rightarrow \max, \quad x \in G_x, \; f_1(x) = C_1.$$

La résolution de ces problèmes nous donne les points $a$ et $b$ (fig. 1.5). En menant par ces points la droite *1*, on obtient une approximation élémentaire de l'ensemble de Pareto.

On améliore cette approximation par la résolution des problèmes III et IV qui nous donne encore deux points, $c$ et $d$, de cet ensemble:

$$\text{III}: \quad f_1(x) \Rightarrow \max, \qquad \text{IV}: \quad f_2(x) \Rightarrow \max,$$
$$x \in G_x, \ f_2 = C_4; \qquad x \in G_x, \ f_1 = C_3.$$

Les valeurs $C_3$ et $C_4$ doivent de nouveau être prises dans l'ensemble permis. La ligne polygonale *2* passant par les points $a$, $c$, $d$, $b$

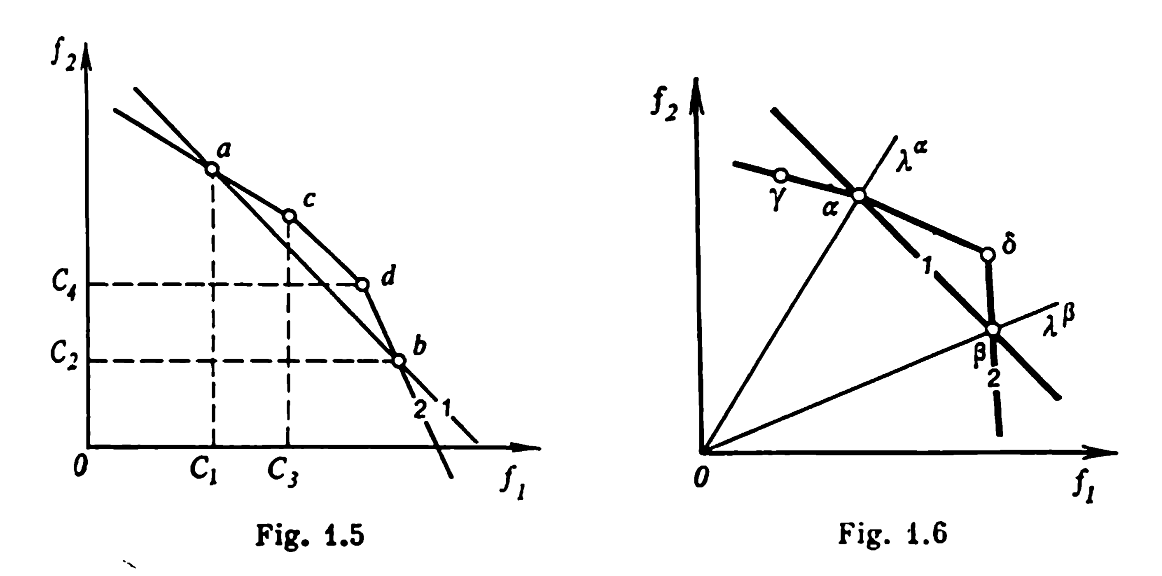



Fig. 1.5

Fig. 1.6

sera l'approximation suivante. Une telle information sur la structure de l'ensemble de Pareto est très souvent suffisante à la résolution de problèmes pratiques. Cette méthode peut être étendue à un plus grand nombre de critères.

Il existe un autre procédé d'approximation de l'ensemble de Pareto. Soient $\lambda_1$ et $\lambda_2$ des nombres $> 0$ tels que

$$\lambda_1 + \lambda_2 = 1. \tag{3.13}$$

Composons le nouveau critère

$$f^\lambda = \lambda_1 f_1(x) + \lambda_2 f_2(x)$$

et résolvons le problème suivant de programmation mathématique:

$$f^\lambda(x) \Rightarrow \max_{x \in G_x}.$$

La solution de ce problème définit un vecteur $\bar{x}$ tel que le point

$$\bar{f}_1 = f_1(\bar{x}), \qquad \bar{f}_2 = f_2(\bar{x})$$

appartienne à l'ensemble de Pareto.

Donc, nous pouvons approcher l'ensemble de Pareto de la manière suivante (fig. 1.6). Résolvons le problème

$$\lambda_1^\alpha f_1(x) + \lambda_2^\alpha f_2(x) \Rightarrow \max_{x \in G_x}, \tag{3.14}$$

où $\lambda_1^\alpha$ et $\lambda_2^\alpha$ satisfont les conditions (3.13). Le problème (3.13), (3.14) définit un vecteur $x_\alpha$ qui nous donne dans le plan $(f_1, f_2)$ un point $\alpha$ de coordonnées

$$f_1 = f_1(x_\alpha), \qquad f_2 = f_2(x_\alpha).$$

Définissons de façon analogue un point $\beta$ et menons par les points $\alpha$ et $\beta$ la droite *1*. Cette droite sera une approximation élémentaire de l'ensemble de Pareto (cf. fig. 1.6). La construction de points $\gamma$ et $\delta$ nous donne la ligne polygonale *2* qui sera l'approximation suivante, etc.

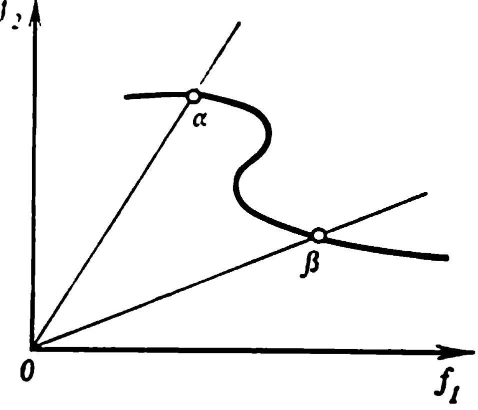



Fig. 1.7

Ces constructions appellent la question suivante: peut-on par ce procédé construire n'importe quel point de l'ensemble de Pareto? En d'autres termes, peut-on à tout point de l'ensemble de Pareto associer un vecteur $\lambda = (\lambda_1, \ldots, \lambda_n)$ satisfaisant les conditions $\sum \lambda_i = 1$, $\lambda_i > 0$, $i = 1, \ldots, n$, et tel que la solution du problème d'optimisation

$$(\lambda, f) = \sum \lambda_i f_i(x) \Rightarrow \max_{x \in G_x}$$

définisse un ensemble de nombres $f_i$ qui soient les coordonnées de ce point? La question reste ouverte dans le cas général. La réponse est positive pour le cas seulement où l'ensemble $G_x$ est un polyèdre et les critères, de la forme

$$f^s = (a^s, x),$$

c'est-à-dire sont aussi des fonctions linéaires (cf. par exemple [26]).

Concluons ce paragraphe par une remarque sur la précision de la méthode d'approximation de l'ensemble de Pareto.

Si l'ensemble de Pareto est convexe, en augmentant le nombre de points définis par l'une des méthodes ci-dessus, on peut construire un polyèdre approchant cet ensemble à n'importe quelle précision. Ceci est illustré par les exemples représentés sur les figures 1.5 et 1.6. Malheureusement la pratique nous confronte à des ensembles de Pareto qui ne sont pas convexes. Ce qui complique énormément le problème de leur approximation. Cette situation est illustrée sur la figure 1.7.

## § 4. Autres types d'indéterminations

a) *Indéterminations naturelles.* Au paragraphe précédent nous avons évoqué les difficultés liées à l'existence d'indétermination des buts. Mais cette indétermination, comme nous l'avons déjà signalé, n'est pas la seule rencontrée par l'analyste.

Le second type d'indétermination sera appelé *indétermination naturelle.* Supposons que nous connaissions notre but, par exemple que nous voulions tracer un itinéraire aérien Moscou — Vladivostok et utiliser les réserves de carburant de telle sorte que le temps de vol $T$ soit minimal. Le temps $T$ ne dépendra pas uniquement de nous, mais aussi des conditions météorologiques sur la ligne. Cette situation est typique et on la décrit ordinairement par un schéma formalisé bien défini. La fonction objectif (par exemple le temps de vol) est mise sous la forme[1]

$$T = f(x, \alpha), \tag{4.1}$$

où $\alpha \in G_\alpha$ est un paramètre (ou une fonction) à priori inconnu et incontrôlable. Le choix de $x$, c'est-à-dire d'une ligne d'action qui minimisera $T$, dépendra de toute évidence essentiellement de $\alpha$.

Donc, quand on parlera d'une indétermination naturelle, on sous-entendra le choix d'une ligne d'action dans le cas où la fonction objectif est donnée, mais pas exactement, c'est-à-dire qu'elle contient un paramètre indéfini. La résolution du problème

$$f(x, \alpha) \Rightarrow \max_x$$

nous permet de définir le vecteur $x$ comme une fonction du seul paramètre $\alpha$:

$$x = x(\alpha). \tag{4.2}$$

Si l'on ne dispose d'aucune information sur l'indétermination $\alpha$, le résultat de l'optimisation de $f(x, \alpha)$ sera arbitraire. Dans les situations réelles, l'information sur le paramètre $\alpha$ est généralement de la nature suivante:

$$\alpha \in G_\alpha$$

où $G_\alpha$ est un ensemble.

Mais une telle information ne suffit pas non plus à assurer une solution unique du problème. La formule (4.2) ne définit qu'une application de l'ensemble des indéterminations naturelles $G_\alpha$ sur l'ensemble $G_x$ que nous appellerons *ensemble d'indétermination du résultat.*

L'ensemble d'indétermination $G_x$ est de toute évidence une très importante caractéristique de l'opération étudiée, mais sa construction implique des calculs complexes. Il existe cependant une approche qui fournit une estimation rigoureuse certes, mais unilatérale: c'est le *principe du meilleur résultat garanti* (ou principe généralisé

du minimax). Expliquons-nous. Comme pour tout $x$

$$\min_{\alpha \in G_\alpha} f(x, \alpha) \leqslant f(x, \alpha), \tag{4.3}$$

on a pour tout $\alpha \in G_\alpha$

$$f^* = \max_x \min_{\alpha \in G_\alpha} f(x, \alpha) \leqslant \max_x f(x, \alpha). \tag{4.4}$$

Le nombre $f^*$ défini par la formule (4.4) s'appelle espérance, et le vecteur $x = x^*$ correspondant, stratégie de prudence au sens que quelle que soit la valeur prise par le paramètre d'indétermination $\alpha$, le vecteur $x = x^*$ fait prendre à la fonction objectif une valeur $\leqslant f^*$. Pour obtenir une stratégie de prudence, il faut résoudre les problèmes d'optimisation suivants:

1) calculer $\min\limits_{\alpha \in G_\alpha} f(x, \alpha)$ pour tout $x$; ceci nous donnera $\alpha = \alpha^*(x)$ et $\hat{f}(x) = f(x, \alpha^*(x))$;

2) calculer $\max\limits_x f(x, \alpha^*(x))$; on obtiendra en définitive $x = x^*$ et $f^*(x^*) = f^*$.

L'espérance peut être considérablement améliorée si l'on sait à l'avance que la valeur du paramètre $\alpha$ sera connue à l'instant de l'« action » (ou de l'« expérience »). Définissons une valeur de la fonction $x = \tilde{x}(\alpha)$ telle que pour tout $\alpha$

$$\max_x f(x, \alpha) = f(\tilde{x}(\alpha), \alpha) = \tilde{f}(\alpha). \tag{4.5}$$

Trouvons ensuite un $\alpha = \tilde{\alpha}$ tel que

$$\min_{\alpha \in G_\alpha} \tilde{f}(\alpha) = \tilde{f}(\tilde{\alpha}) = \min_{\alpha \in G_\alpha} \max_x f(x, \alpha).$$

Comme

$$\min_{\alpha \in G_\alpha} \max_x f(x, \alpha) \geqslant \max_x \min_{\alpha \in G_\alpha} f(x, \alpha) = f^*,$$

le fait qu'au moment de la prise de décision l'on connaîtra la valeur du facteur indéterminé, par exemple la situation météorologique, nous permet d'obtenir une nouvelle espérance plus « parfaite ». Mais dans ce cas la stratégie de prudence ne sera pas le vecteur $x = x^*$, mais une fonction

$$x = \tilde{x}(\alpha).$$

Le choix d'une stratégie de prudence est un moyen rationnel de prise de décision. L'usage de cette stratégie nous met à l'abri de tout aléas: quels que soient les facteurs incontrôlables, on s'assure une

valeur de la fonction objectif $\leqslant f^*$. Mais on garde toujours la possibilité d'améliorer ce résultat. Il faut à cet effet prendre une décision comportant un certain risque: en donnant une certaine valeur au facteur indéterminé, on peut tout aussi bien obtenir une plus grande qu'une plus petite valeur de la fonction objectif. Considérons plus en détail une situation à risque. On distingue généralement deux cas extrêmes: le choix est effectué plusieurs fois et le choix a lieu une seule fois. Dans les deux cas, on admet que $\alpha$ est une variable aléatoire dont la loi de probabilité est connue. On voit qu'entre ces deux cas la différence n'est pas si importante.

La variable $\alpha$ étant aléatoire, il en sera de même de la fonction $f(x, \alpha)$. Donc, dans le cas où l'on a affaire à des opérations itératives, on gagne à remplacer le problème donné par un problème de probabilité. Cette situation n'est pas nouvelle, puisqu'on y a été confronté au § 2. Pour estimation de la stratégie choisie, on peut maintenant retenir le maximum de l'espérance mathématique

$$f_1 = \max_x \overline{f(x, \alpha)}.$$

Mais la substitution du problème $\overline{f(x, \alpha)} \Rightarrow \max$ au problème $f(x, \alpha) \Rightarrow \max$ n'est pas le seul moyen de passer à la position stochastique. On a vu qu'on pouvait prendre pour estimation la quantité

$$f_2 = \overline{\max_x f(x, \alpha)} = \overline{f(\hat{x}(\alpha), \alpha)},$$

où $x = \hat{x}(\alpha)$ est la meilleure stratégie pour $\alpha$ connu.

Il peut exister d'autres critères. Soit par exemple $\bar{\alpha}$ la valeur moyenne de la variable aléatoire $\alpha$; il lui correspondra une fonction $f_3 = f(x, \bar{\alpha})$ dont le maximum peut également être adopté pour estimation.

Le remplacement de l'un de ces problèmes par n'importe quel autre est un acte non formel, car ces problèmes sont fondamentalement différents. D'une façon générale, quel que soit le critère formulé dans cette situation, son choix ne sera pas une opération mathématique rigoureuse. Il traduira notre hypothèse, savoir que la détermination des paramètres du système à l'aide de la nouvelle règle introduite lui garantira la qualité désirée. Cette situation est la conséquence du fait évident que la condition initiale

$$f(x, \alpha) \Rightarrow \max \tag{4.6}$$

équivaut au fond à une infinité de critères généralement distincts. En effet, supposons que le paramètre $\alpha$ ne prend que des valeurs discrètes $\alpha_1, \alpha_2, \ldots$; alors la condition (4.6) équivaut à la maxi-

misation de l'ensemble des critères:

$$\begin{aligned} &f(x, \alpha_1) \Rightarrow \max, \\ &f(x, \alpha_2) \Rightarrow \max, \\ &\dots\dots\dots\dots \\ &f(x, \alpha_n) \Rightarrow \max, \\ &\dots\dots\dots\dots \end{aligned} \tag{4.7}$$

Donc, le problème décisionnel dans le cas où le paramètre caractérisant les indéterminations naturelles est aléatoire, présente beaucoup d'affinités avec le problème décisionnel dans le cas d'une indétermination des buts. Au paragraphe précédent, on a vu que dans cette situation il faut introduire une hypothèse supplémentaire: réaliser une réduction des critères. Or la réduction des critères est toujours un acte non formel et n'importe lequel des critères $f_1$, $f_2$ et $f_3$ à l'aide desquels nous avons cru possible de choisir une stratégie ne sera qu'une hypothèse. Cette assertion vaut aussi bien pour les opérations itératives que non itératives. Mais la réduction des critères qui permet de passer à la position stochastique est logique dans les opérations itératives. Ce qui vient d'être dit signifie que ce passage se justifie intuitivement (par l'expérience des gens), bien qu'il traduise l'idée subjective que l'on se fait du but de l'opération. Une telle formalisation ne contredit pas les objectifs poursuivis, elle les reflète assez bien dans le cas d'opérations itératives. La situation est différente si le choix $x$ ne se fait qu'une seule fois. Dans ce cas l'information que $\alpha$ est une variable aléatoire dont les caractéristiques statistiques sont connues n'est pratiquement d'aucune utilité.

Et pourtant l'on sait déjà par l'analyse de l'exemple de construction des silos que l'interprétation stochastique peut être utilisée dans le cas d'opérations non itératives, car elle fournit une des éventuelles réductions des critères.

Signalons qu'on a déjà eu affaire à une situation analogue lors de l'analyse des contraintes. Au § 1 nous avons indiqué un exemple de contraintes physiques

$$\sum_i q_i x_i \leqslant Q, \tag{4.8}$$

où $Q$ est la capacité du barrage. La quantité $Q$ est aléatoire, vu qu'elle dépend des conditions météorologiques. Supposons que $Q$ ne prend que des valeurs discrètes. La con ition (4.8) équivaut alors aux inégalités suivantes:

$$\begin{aligned} &\sum_i q_i x_i \leqslant Q_1, \\ &\sum_i q_i x_i \leqslant Q_2, \\ &\dots\dots \end{aligned}$$

La condition (4.8) est physique et l'on est contraint de la respecter: en effet, on ne peut fournir à la terre plus d'eau qu'on n'en a à sa disposition. En choisissant $x_i$ on ne connaît pas à l'avance la valeur de $Q$. Mais pour prendre une décision, c'est-à-dire pour choisir $x_i$, nous devons fixer d'une manière quelconque la valeur de $Q$. La désignation de $Q$ ou le passage à la description stochastique sera toujours une nouvelle hypothèse.

Comment accorder la décision prise avec la réalité consécutive au choix de la quantité $x_i$? En effet, ceci n'est pas aussi simple, puisque l'on ne peut pas violer la contrainte physique. On est donc mis en présence d'une nouvelle opération: redistribuer l'eau de manière à maximiser la fonction objectif dans les nouvelles conditions (après avoir obtenu une nouvelle information). Mais rappelons que ce sera un nouveau problème.

En attribuant telle ou telle valeur au paramètre aléatoire $Q$ ou en adoptant pour critère une caractéristique stochastique du processus, on ne pourra jamais indiquer à l'avance le résultat de l'opération. Certes, on dispose toujours d'une stratégie de prudence. Désignons par exemple par $Q_*$ le minimum de $Q$; alors en choisissant $x_i$ telles que

$$\sum_i q_i x_i \leqslant Q_*,$$

on réalisera nécessairement la condition (4.8) et l'on aura la majoration garantie, c'est-à-dire que l'on pourra garantir la récolte minimale.

Dans tous les autres cas, on aura affaire à de nouvelles hypothèses, c'est-à-dire à un risque dont l'acceptation dépend entièrement de l'analyste ou du responsable.

b) *Partenaire actif*. Passons maintenant à la description des indéterminations liées à l'existence de partenaires ou d'adversaires actifs dont on ne peut entièrement contrôler le comportement.

La théorie de la recherche opérationnelle réserve une place particulière à l'étude des situations oligopoles, c'est-à-dire confrontant plusieurs personnes poursuivant chacune un objectif personnel

$$f_i(x_1, x_2, \ldots, x_n) \Rightarrow \max_{x_i}$$

avec des moyens décrits par un vecteur $x_i$, $x_i \in G_i$.

A noter que formellement cette situation comprend un problème à plusieurs critères impliquant la quête d'un vecteur $x$ maximisant les critères $f_i(x)$. En effet, si l'on identifie l'objectif de chaque personne à son critère $f_i(x_i)$ et que l'on décrive l'ensemble $G_i$ à l'aide de la condition

$$x_1 = x_2 = \ldots = x_n,$$

alors on obtient le cas particulier d'un problème à plusieurs partenaires actifs. Il est certes évident que le cas général d'une situation oligopole est bien plus complexe et son analyse implique toute une série d'hypothèses spécifiques. Illustrons ceci sur l'exemple d'une situation duopole.

Supposons donc que deux personnes $A$ et $B$ ayant la possibilité de choisir des vecteurs $x$ et $y$ tentent de réaliser leurs propres objectifs représentés sous la forme

$$f(x, y) \Rightarrow \max, \quad \varphi(x, y) \Rightarrow \max, \quad x \in X, \quad y \in Y.$$

On peut avoir en particulier $f = -\varphi$; une telle situation sera dite antagoniste.

REMARQUE. Les situations antagonistes ont fait l'objet d'innombrables recherches, notamment en théorie des jeux, discipline mathématique qui a vu le jour grâce aux travaux du mathématicien français E. Borel sur les jeux de société. Les situations antagonistes pures sont dans un certain sens dégénérées. Un exemple type est le conflit dans lequel les intérêts des partenaires ou des adversaires qui sans être identiques ne sont pas strictement opposés.

On appellera *conflit* le cas général d'incompatibilité des intérêts (buts) des partenaires. Dans l'étude des situations de conflit, il est commode d'identifier l'analyste à l'une des personnes. On conviendra par exemple de dire « nous » pour la personne $A$. Cela se conçoit car l'analyse est toujours effectuée du point de vue des intérêts d'une personne.

Etant donné que le résultat de notre choix dépend de celui de la personne $B$, nous devons adopter telle ou telle hypothèse sur son comportement, comportement qui à son tour dépendra du caractère de l'information dont dispose la personne $B$. Plusieurs hypothèses (cas) sont possibles.

$\alpha$) Aucune des personnes n'est informée sur la politique de l'autre. Dans ce cas nous pouvons trouver l'espérance. Pour la personne $A$ elle est donnée par la formule

$$f^* = \max_{x \in X} \min_{y \in Y} f(x, y), \tag{4.9}$$

pour la personne $B$, par

$$\varphi^* = \max_{y \in Y} \min_{x \in X} \varphi(x, y). \tag{4.9'}$$

La résolution des problèmes (4.9) et (4.9′) nous donne des vecteurs $x^*$ et $y^*$ tels que si l'on fait $x = x^*$, alors quelles que soient les conditions (quel que soit $y \in Y$), la valeur prise par la fonction objectif $f(x, y) \geqslant f^*$.

Plusieurs risques peuvent être pris dans cette situation. On peut par exemple supposer que l'autre personne utilise la stratégie de

prudence $y = y^*$. Dans ce cas notre choix sera différent:

$$f(x, y^*) \Rightarrow \max_{x \in X}.$$

On définit le vecteur $x = x^{**}$ et la valeur respective de la fonction $f = f^{**}$ [$f^{**} \geqslant f^*$). Mais si l'adversaire (le partenaire) procède à un autre choix, par exemple $y = y^{**}$, alors il est possible que $f(x^{**}, y^{**}) < f^*$. Mais un risque est un risque: nous avons émis une hypothèse et si elle est fausse, le résultat peut être différent.

REMARQUE. Les raisonnements de la nature suivante offrent de grandes possibilités dans la formation des hypothèses de risque: la personne $A$ peut par exemple supposer que son adversaire, la personne $B$, supposera qu'elle a choisi sa stratégie de prudence et dans cette hypothèse optera lui aussi pour la sienne, etc.

β) Supposons qu'au moment d'agir (de choisir $x$) nous, c'est-à-dire la personne $A$, sachions la valeur $y$ retenue par $B$.

Nous devons chercher notre stratégie, c'est-à-dire $x$, sous forme d'une fonction $x = x(y)$. Nous pouvons la déterminer réellement. Nous devons à cet effet résoudre le problème d'optimisation

$$f(x, y) \Rightarrow \max_{x \in X}. \qquad (4.10)$$

La condition (4.10) définit la stratégie cherchée $x = \hat{x}(y)$.

Nous pouvons pour ce cas aussi calculer l'espérance $\hat{f}$; il sera différent de $f^*$:

$$\hat{f} = \min_{y \in Y} \max_{x \in X} f(x, y),$$

et dans tous les cas $f^* \leqslant \hat{f}$.

A noter qu'en choisissant notre stratégie (le vecteur $x$) nous ne pouvons en aucun cas influencer le choix de l'adversaire.

γ) Supposons maintenant qu'au moment de prendre sa décision la personne $B$ soit au courant de notre choix; nous sommes par exemple dans l'obligation de l'en informer. Dans ce cas, nous pouvons influencer son choix. En effet, si nous connaissons la fonction objectif de la personne $B$, nous admettrons naturellement que celle-ci définira sa politique à partir de la condition

$$\varphi(x, y) \Rightarrow \max_{y \in Y}. \qquad (4.11)$$

La résolution du problème (4.11) nous donnera la réponse de la personne $B$ à notre choix, qui, d'après notre hypothèse, est sa stratégie optimale:

$$y = \hat{y}(x). \qquad (4.12)$$

Nous pouvons maintenant procéder au choix du vecteur $x$. En effet, en portant (4.12) dans l'expression de la fonction objectif

$f(x, y)$, on obtient

$$f(x, \hat{y}(x)) = F(x), \tag{4.13}$$

et l'on peut maintenant déduire notre stratégie à partir de la condition

$$F(x) \Rightarrow \max_{x \in X}. \tag{4.14}$$

Donc, si l'on sait que la personne $B$ est informée de notre choix et qu'elle y ripostera par sa stratégie optimale, on peut alors influer sur sa politique de telle sorte qu'elle corresponde au maximum à notre propre objectif.

Si le maximum (4.11) est réalisé non pas en un point mais sur un ensemble $M(x)$, alors l'espérance de $A$ est donnée par le calcul du minimum sur cet ensemble et la meilleure espérance de $A$, par le calcul du maximum sur $x \in X$, c'est-à-dire

$$f^{***} = \max_{x \in X} \min_{y \in M(x)} f(x, y).$$

Cette situation est assez fréquente en pratique et on lui trouve immédiatement une interprétation économique. Ainsi le vecteur $x$ peut décrire des ressources, la fonction (4.12), la fonction de production qui décrit le meilleur moyen d'utilisation des ressources par la personne $B$. La personne $B$ dispose donc de la quantité de ressources qui fera le mieux correspondre ses activités à nos objectifs.

δ) *Problème de pénalisation et de récompense.* Au numéro β) nous avons vu une situation dans laquelle notre stratégie revêtait un caractère de synthèse:

$$x = \hat{x}(y). \tag{4.15}$$

Supposons maintenant que nous avons la possibilité non seulement de choisir la stratégie (4.15) mais aussi de la communiquer à la personne $B$. Il se trouve que cette information nous permet de déterminer la riposte aux actes de la personne $B$, qui corresponde le mieux à nos objectifs, et d'influencer la politique de la personne $B$.

En effet, l'hypothèse la plus naturelle qu'on puisse faire sur le comportement de la personne $B$ est qu'elle essayera de tirer parti du fait qu'elle connaît notre stratégie $\hat{x}(y)$ pour optimiser sa propre fonction objectif, qui s'écrit ici

$$\varphi(x, y) = \varphi(\hat{x}(y), y). \tag{4.16}$$

Le problème de la personne $B$ est de trouver maintenant un $y$ qui maximise la fonction $\varphi^*(y) = \varphi(\hat{x}(y), y)$. La résolution de ce problème nous donne $y$ comme un opérateur de notre stratégie $\hat{x}(y)$:

$$y = \hat{y}[\hat{x}(y)]. \tag{4.17}$$

L'expression (4.17) signifie qu'à toute stratégie $\hat{x}(y)$ est associé un vecteur $y$ bien défini: l'expression (4.17) est une application de l'ensemble de nos stratégies sur l'ensemble des choix de la personne $B$. Une fois qu'on connaît la réponse de $B$ à notre stratégie $\hat{x}(y)$, on peut s'occuper du choix de notre stratégie optimale. Il suffit pour cela de résoudre le problème

$$f(\hat{x}(y), \hat{y}[\hat{x}(y)]) \Rightarrow \max_{\hat{x}(y)} \qquad (4.18)$$

C'est un problème spécial d'optimisation dont la résolution nous donne la stratégie optimale $\hat{x}(y)$ et la valeur correspondante de la fonction objectif.

Si en maximisant la fonction (4.16) on trouve plusieurs vecteurs $y$ et si l'ensemble des vecteurs est $M(\hat{x})$, alors la meilleure espérance de la personne $A$ sera

$$f^{****} = \max_{\hat{x} \in X_1} \min_{y \in M(\hat{x})} f(\hat{x}(y), y),$$

où $\hat{x} = \hat{x}(y)$ et de plus $\hat{x} \in X_1$, $X_1$ étant l'ensemble des fonctions à valeurs dans $X$.

Cette situation admet une interprétation économique simple. Elle décrit le schéma de fonctionnement d'organes économiques dont l'un a la possibilité d'agir sur le comportement de l'autre par le biais de pénalités ou de récompenses décrites par la fonction $\hat{x}(y)$.

On pourrait envisager beaucoup d'autres situations semblables *). Nous aurons divers cas d'interaction des partenaires selon le caractère de l'information mise entre les mains des personnes.

Nous avons envisagé ici des situations que l'on pourrait qualifier d'idéales. Nous avons admis que les deux partenaires non seulement connaissaient leurs propres objectifs mais étaient complètement informés sur ceux de l'autre. De tels cas sont assez rares. La situation suivante est plus fréquente. Nous ne savons jamais avec certitude les buts de nos partenaires ou de nos adversaires, et nos adversaires, c'est-à-dire les autres personnes du système, connaissent vaguement nos buts. La personne qui analyse la situation conflictuelle doit toujours en tenir compte. Une idée erronée de la situation réelle des personnes complique la formation des hypothèses sur le comportement, hypothèses sans lesquelles il est impossible de prendre une décision plus ou moins adéquate. Pour construire des hypo-

---

*) D'intéressantes et importantes recherches ont été consacrées ces dernières années à l'analyse de ces questions. Le lecteur désireux d'approfondir ses connaissances en la matière peut consulter les ouvrages [4, 5, 72].

thèses sur le comportement des autres personnes, nous devons en général formuler des hypothèses sur leur information.

c) *Situations d'équilibre.* Nous avons vu au paragraphe précédent que l'analyse des indéterminations s'appuie sur de nombreuses hypothèses qui nous ont permis de dégager de l'ensemble des solutions un sous-ensemble de solutions admissibles. Cette analyse nous conduit à exclure les « mauvaises » solutions qui ne peuvent aucunement prétendre être optimales. En examinant le problème d'indétermination des buts nous nous sommes longuement attardés sur le principe de Pareto qui est l'un des plus importants principes de sélection des solutions rationnelles. L'analyse parétienne définit les conditions que doit nécessairement remplir tout compromis raisonnable. Le principe de Pareto reste en vigueur pour l'analyse des situations conflictuelles oligopoles. L'un des problèmes les plus importants qui se pose dans ces situations est le problème des solutions collégiales, de la formation collégiale du compromis. Il est évident que dans une telle situation il faut rejeter toutes les solutions (les décisions) qui peuvent être remplacées par d'autres faisant prendre de plus grandes valeurs aux fonctions objectifs de toutes les personnes simultanément ou d'une partie d'entre elles, mais sans diminuer les valeurs des fonctions objectifs des autres. En un mot, on discutera la validité des seules solutions collégiales appartenant à l'ensemble de Pareto. Ces variantes (qui seront appelées effectives) ont la propriété d'améliorer la valeur de la fonction objectif d'une personne au détriment des autres.

Pour analyser les situations oligopoles, il existe des principes autres que celui de Pareto. Arrêtons-nous sur l'un d'eux, appelé *principe de stabilité* ou *d'équilibre.* Ce principe a été établi en théorie des jeux dont l'objet est l'analyse des conflits entre deux personnes. Considérons une situation duopole et supposons que le but du joueur $A$ est de maximiser une fonction $f(x, y)$ à l'aide d'un vecteur $x$ qu'il peut choisir dans un ensemble $X$. En d'autres termes, le but du joueur $A$ s'écrit:

$$f(x, y) \Rightarrow \max_{x \in X}. \tag{4.19}$$

La situation étant antagoniste, le joueur $B$ poursuivra un but diamétralement opposé, c'est-à-dire qu'il tentera par tous les moyens de faire en sorte que

$$f(x, y) \Rightarrow \min_{y \in Y}. \tag{4.20}$$

A noter que dans cette situation il est question non pas d'extrémums locaux mais d'extrémums absolus sur les ensembles respectifs et la tâche principale consiste à formuler des recommandations simultanément pour les deux joueurs quant à leur procédure de choix. Dans les positions classiques on parle toujours d'une analyse « abso-

lue » ou « impartiale » réalisée par une tierce personne qui ne poursuit aucun but et qui a accès à n'importe quelle information.

Quelles recommandations peut-on prodiguer ici?

Supposons que la fonction $f(x, y)$ présente un col (fig. 1.8) sur le produit direct des ensembles $X$ et $Y$. En ce point on a l'égalité évidente qui est la définition du point col:

$$f^* = \max_{x\in X} \min_{y\in Y} f(x, y) = \min_{y\in Y} \max_{x\in X} f(x, y). \tag{4.21}$$

Soient $(x^*, y^*)$ les coordonnées du point col. Il est évident qu'aucun des joueurs n'a intérêt à prendre pour stratégie une stratégie autre

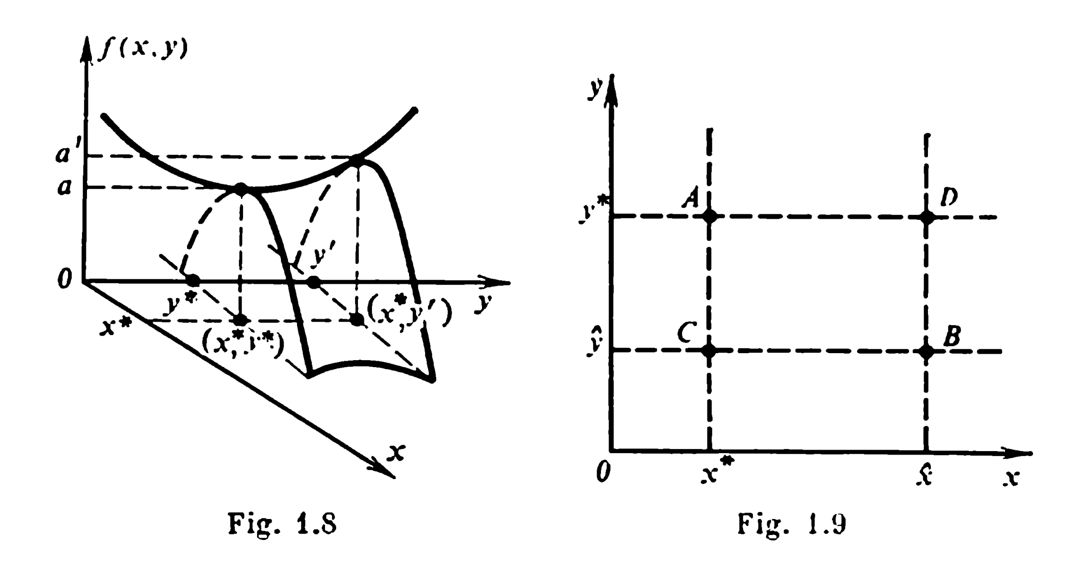



Fig. 1.8 Fig. 1.9

que $x^*$ ou $y^*$. Supposons par exemple que le joueur $B$ ait choisi une stratégie $y = y'$ à la place de $y = y^*$. On voit sur la figure que le minimum de sa fonction objectif sous réserve d'une politique raisonnable de l'adversaire $A$ (c'est-à-dire dans le cas où ce dernier choisit $x = x^*$) sera $a' > a$. En ce sens le col est un point de choix stable. Comme il est question d'extrémum absolu, cette situation est valable dans le cas général de plusieurs cols.

Supposons par exemple que la fonction $f(x, y)$ présente deux cols $(x^*, y^*)$ et $(\hat{x}, \hat{y})$ en lesquels, ainsi qu'il résulte de (4.21), on a

$$f(x^*, y^*) = f(\hat{x}, \hat{y}) = f^*. \tag{4.22}$$

A noter que la propriété (4.22) n'est pas la seule propriété remarquable des points cols. En effet, la fonction $f(x, y)$ présente des cols autres que $(x^*, y^*)$ et $(\hat{x}, \hat{y})$, en l'occurrence les points $(x^*, \hat{y})$ et $(\hat{x}, y^*)$ en lesquels

$$f(x^*, \hat{y}) = f(\hat{x}, y^*) = f^*.$$

Prouvons cette relation. Soit $A$ et $B$ des points cols (fig. 1.9) de coordonnées respectives $(x^*, y^*)$ et $(\hat{x}, \hat{y})$. Considérons le point $C$ de coordonnées $(x^*, \hat{y})$ (les raisonnements sont les mêmes pour le point $D$ $(\hat{x}, y^*)$). Les points $(\hat{x}, \hat{y})$ et $(x^*, y^*)$ étant des cols, on a

$$f(x, \hat{y}) \leqslant f(\hat{x}, \hat{y}) \leqslant f(\hat{x}, y), \tag{4.23}$$

$$f(x, y^*) \leqslant f(x^*, y^*) \leqslant f(x^*, y), \tag{4.24}$$

$\forall x \in X$ et $\forall y \in Y$, d'où, (4.21) et (4.22) aidant, on obtient les inégalités

$$f^* = f(x^*, y^*) \leqslant f(x^*, \hat{y}) \leqslant f(\hat{x}, \hat{y}) = f^*.$$

D'où il s'ensuit

$$f(x^*, \hat{y}) = f^*.$$

De (4.23) et (4.24) il vient par ailleurs

$$f(x^*, \hat{y}) = f(x^*, y^*) \leqslant f(x^*, y)$$

et

$$f(x^*, \hat{y}) = f(\hat{x}, \hat{y}) \geqslant f(x, \hat{y}).$$

Ces relations expriment que $(x^*, \hat{y})$ est un col.

Si donc l'on a deux cols ($A$ et $B$), on trouve alors immédiatement encore deux autres ($C$ et $D$) et la fonction $f(x, y)$ prend des valeurs égales en ces points.

L'on comprend maintenant le rôle des points cols en théorie des conflits. S'il existe plusieurs cols, chaque joueur peut utiliser indifféremment la stratégie $y^*$ ou $\hat{y}$. N'importe laquelle de ces stratégies garantit au joueur $B$ la même valeur ($\geqslant f^*$) de la fonction objectif. Si l'autre joueur choisit l'une des stratégies $x^*$ ou $\hat{x}$, alors sa fonction objectif prendra la valeur $f^*$ ; cette valeur sera également prise par la fonction objectif du joueur $A$.

Donc, l'existence d'un col nous permet bien de parler d'une solution optimale du point de vue des deux joueurs et la prise d'une décision se ramène seulement à la détermination du maximin. Ceci explique que pendant de longues années en théorie classique des jeux les efforts des mathématiciens étudiant les situations de conflit aient été concentrés sur des problèmes se ramenant à l'étude de points cols (ou bien à une modification de la position du problème qui conduise en fin de compte à l'analyse d'une situation d'équilibre, c'est-à-dire à l'analyse de points cols).

L'importance exceptionnelle des situations d'équilibre dans les conflits a inspiré naturellement des tentatives d'extension de la notion d'équilibre au cas général de plusieurs joueurs. Soit un sys-

tème de $N$ joueurs visant par le choix de leurs stratégies $x_i \in X_i$ à maximiser leurs fonctions objectifs $f_i$. Les valeurs de $f_i$ dépendront généralement non seulement du choix du joueur $i$ mais aussi de celui des autres joueurs, c'est-à-dire

$$f_i = f_i(x_1, \quad x_2, \ldots, x_{i-1}, \quad x_i, x_{i+1}, \ldots, x_N).$$

On dira qu'un point (choix) $\hat{x} = \{\hat{x}_1, \ldots, \hat{x}_N\}$ est une *situation d'équilibre* si pour tout $i$ on a

$$\max_{x_i} f_i(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_{i-1}, x_i, \hat{x}_{i+1}, \ldots, \hat{x}_N) = f_i(\hat{x}_1, \ldots, \hat{x}_i, \ldots, \hat{x}_N). \tag{4.25}$$

Les points d'équilibre seront naturellement appelés stables, car si le joueur $i$ s'écarte de sa valeur d'équilibre, c'est-à-dire choisit une stratégie distincte de $\hat{x}_i$, alors il perdra au cas où les autres joueurs maintiendraient leurs choix, puisque

$$f_i(\hat{x}_1, \ldots, \hat{x}_{i-1}, x_i, \hat{x}_{i+1}, \ldots, \hat{x}_N) \leqslant \leqslant f_i(\hat{x}_1, \ldots, \hat{x}_i, \ldots, \hat{x}_N). \tag{4.26}$$

Ceci est à l'origine du principe de stabilité (ou principe de Nash, du nom de son auteur) qui affirme que la stratégie rationnelle doit être choisie dans l'ensemble des points d'équilibre, c'est-à-dire des points vérifiant la condition (4.25) (ou (4.26)).

Voyons dans quelle mesure ce principe est universel et jusqu'à quel point on peut l'utiliser pour estimer la qualité des choix. Est-il si évident d'exclure l'ensemble des choix ne vérifiant pas le principe de Nash? Signalons tout d'abord que dans le cas de plusieurs joueurs poursuivant chacun des objectifs propres distincts, il faut toujours envisager un compromis: lorsque des décisions collectives sont à prendre, chacun des joueurs doit dans une certaine mesure faire des concessions. C'est pourquoi la condition de stabilité est une très importante propriété de compromis.

Si tous les joueurs conviennent de s'en tenir au choix $x_i = \hat{x}_i$, alors celui d'entre eux qui viole cet accord en fait le premier les frais: la stabilité est une garantie notoire contre la violation de la convention. Nous disons bien la convention, c'est-à-dire la décision collective. C'est là que réside la différence de principe entre le cas général et le cas d'un conflit en présence de col que nous venons tout juste d'examiner et où la notion de compromis n'a pas de sens.

En effet, dans le cas d'un conflit entre deux personnes il n'existe aucune convention sur le choix de la stratégie optimale, puisque le seul comportement rationnel est celui qui consiste à prendre la stratégie d'équilibre (si elle existe).

La situation est fondamentalement différente dans le cas de plusieurs joueurs. Considérons l'exemple notoire de Yu. Hermeyer. Supposons que les fonctions objectifs des joueurs sont de la forme

$$f_i(x_1, \ldots, x_N) = x_i + \sum_{i \neq j} (1 - x_j), \tag{4.27}$$

les « actions » des joueurs (les quantités scalaires $x_i$) étant soumises aux contraintes

$$0 \leqslant x_i \leqslant 1, \qquad i = 1, \ldots, N. \tag{4.28}$$

Le point d'équilibre est de toute évidence

$$\hat{x}_i = 1, \qquad i = 1, 2, \ldots, N, \tag{4.29}$$

et $f_i(\hat{x}_1, \ldots, \hat{x}_N) = \hat{f}_i = 1$. C'est un point stable. En effet, si tous les joueurs à l'exception du joueur $i$ s'en tiennent à la stratégie (4.29) et le joueur $i$, à la stratégie

$$x_i = 1 - \delta$$

(en vertu de la condition (4.28) le joueur $i$ ne peut que diminuer la quantité $x_i$ par rapport à sa valeur d'équilibre), alors la valeur de sa fonction objectif satisfera la condition

$$f_i = 1 - \delta < \hat{f}_i,$$

c'est-à-dire sera $<$ à la valeur que lui aurait assuré le choix de la stratégie d'équilibre $x_i = \hat{x}_i = 1$. Mais la stratégie d'équilibre ne sera pas optimale dans ce cas. Il existe une infinité de stratégies qui font prendre aux fonctions objectifs des valeurs $>$ à la valeur d'équilibre, ce pour tous les joueurs simultanément. Si l'on admet par exemple que $x_i = 0$, on trouve pour un tel choix

$$f_i = N - 1 > 1 \qquad \text{pour } N > 2.$$

Le point $x_i = 0$, $i = 1, \ldots, N$, n'est visiblement pas stable. En effet, supposons que tous les joueurs à l'exception de $i$ choisissent les stratégies $x_j = 0$ et le joueur $i$, la stratégie $x_i = 1$. Sa fonction objectif $f_i$ prendra alors la valeur $N$ et non pas $N - 1$. Pour ce qui est des autres joueurs, ils verront les valeurs de leurs fonctions objectifs diminuer du fait que le joueur $i$ renonce à sa stratégie gagnante $x_i = 0$. Dans ce cas leur « gain » sera

$$f_j = N - 2, \qquad j \neq i.$$

Résumons-nous. La stabilité du choix est une très importante propriété dans le cas de plusieurs joueurs ($N > 2$). Mais stable ne veut pas dire efficace, c'est-à-dire que ce choix n'appartient pas forcément à l'ensemble de Pareto. Donc, si tous les partenaires

(joueurs) prennent indépendamment une décision, il y a peu de chance qu'elle soit stable.

Par conséquent le principe de stabilité (principe de Nash) ne peut probablement pas être considéré comme un principe de choix d'une décision. La situation est différente dans le cas d'une décision prise d'un commun accord par tous les joueurs. Mais ici aussi un élément de doute subsiste toujours: une partie des joueurs peut s'entendre sur une autre stratégie (par exemple, une stratégie appartenant à l'ensemble de Pareto). Cette partie s'assurera alors de meilleurs résultats que les autres.

En fait, le seul cas où la condition de stabilité peut être considérée comme un principe d'élimination des variantes non concurrentielles est celui où les points stables sont à la fois des points de l'ensemble de Pareto. Ces systèmes, même s'ils sont très rares, n'en possèdent pas moins (comme nous le verrons) une grande importance pratique. Les situations dans lesquelles les choix efficaces sont instables et les stables, non efficaces, sont plus fréquentes. Les cas où les points stables appartiennent à la fois à l'ensemble de Pareto correspondent toujours à d'importants problèmes pratiques. Ceci explique que l'étude des systèmes dans lesquels les points stables appartiennent à l'ensemble de Pareto constitue un important chapitre de la théorie de la recherche opérationnelle.

## § 5. Commentaire final

Nous venons de jeter un coup d'œil rapide sur certains problèmes de la recherche opérationnelle. Les grandes lignes de cette discipline sont les suivantes:

1) description mathématique (élaboration d'un modèle d'opération);

2) analyse des indéterminations et formalisation de la notion de but (construction de la fonction objectif et du critère);

3) résolution des problèmes d'optimisation et des autres problèmes.

Ce schéma est assez conventionnel, car les divisions indiquées s'imbriquent lors de l'étude d'une opération concrète.

La construction d'un modèle de l'opération, construction qui est le point de départ de toute recherche, implique une profonde compréhension des traits spécifiques du processus et de l'appareil mathématique mis entre les mains de l'analyste. A cette étape il est question de l'aspect « physique » du processus. Nous n'avons pas encore évoqué le but de la recherche, mais celui-ci figure implicitement: les modèles économiques destinés au choix d'une politique de développement des ressources énergétiques d'une région ou au développement de systèmes d'irrigation de la même région seront différents, bien que le développement des ressources énergétiques ne

puisse être conçu séparément de celui de la production agricole ou de toute autre production. Mais les divers blocs ne seront pas tous décrits avec le même détail. Dans un modèle économique destiné au choix des politiques de développement des ressources énergétiques d'une région, l'agriculture, par exemple, sera décrite en indices fortement agrégés (intégraux), alors que toutes les particularités de la production énergétique le seront avec force détails. Dans un modèle simulant un système d'irrigation, la production agricole doit être décrite en détail, mais les ressources énergétiques par contre ne figureront dans ce modèle que par leurs indices agrégats. Ainsi la finalité du modèle exerce une certaine influence sur la procédure initiale de son élaboration.

L'étude de l'information mise entre les mains de l'analyste constitue un maillon important de la recherche. L'analyste doit en effet juger dans quelle mesure cette information est conforme au modèle construit et éventuellement modifier ce dernier en introduisant de nouvelles caractéristiques, etc.

REMARQUE. Selon un point de vue assez répandu, l'information utilisée constitue un facteur décisif dans la formation du modèle. Je m'élève catégoriquement contre de tels jugements. L'information « traditionnelle » utilisée par les projeteurs ou les économistes a été recueillie en fonction des techniques « traditionnelles » de l'analyse, techniques qui ne tenaient pas compte des performances des ordinateurs. Donc, l'étude de l'information doit généralement s'achever par son assujettissement à de nouvelles conditions. Cela étant, l'analyste doit toujours rester réaliste, car toute exigence superflue ne peut que nuire.

Un aspect de l'étude de la correspondance du modèle et de l'information est lié au fait que cette dernière risque d'être inexacte ou pas suffisamment exacte pour des raisons indépendantes de notre volonté. Dans ces conditions, toute description détaillée du modèle (toute tentative de réaliser un modèle « presque conforme » au processus réel) non seulement est inutile, mais est tout simplement nuisible.

L'autre groupe de problèmes est constitué par la formation du critère et des hypothèses pour lever les indéterminations. Il faut commencer par le recensement des critères (indices) et des possibilités limites. Il est fondamental d'étudier l'ensemble de Pareto pour les plus « importants » critères. On arrive finalement à une réduction des critères. Vu qu'il existe plusieurs variantes de réductions, une analyse et une comparaison des résultats s'imposent. Il va de soi qu'il faut éliminer les fonctions critères qui présentent un maximum en « pic ». Expliquons cette assertion sur l'exemple d'un problème de choix des paramètres du système technique projeté.

Les modèles réels et leurs propriétés doivent être stables pour de petites variations des caractéristiques du projet, car la réalisation de ce dernier donnera lieu à d'inévitables écarts par rapport au modèle théorique et ces écarts ne doivent pas trop altérer le modèle

réel. En d'autres termes, la situation de la figure 1.10, *b* est inacceptable. Il faut toujours essayer de « travailler » avec un critère dont la dépendance par rapport aux paramètres de construction a la forme représentée sur la figure 1.10, *a*.

L'appréciation de l'espérance (stratégie minimax) occupe une très importante place en analyse. Le calcul de la stratégie minimax

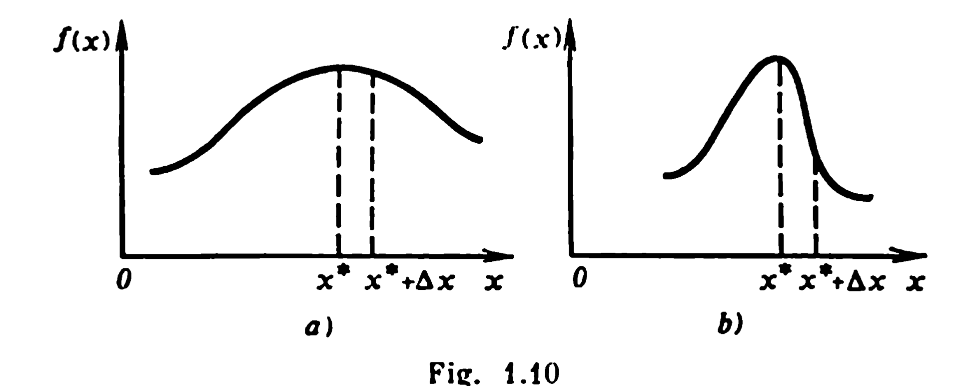



Fig. 1.10

est la pierre angulaire de l'étude de toute opération, le seul résultat objectif de l'analyse qui ne dépende pas de l'information.

REMARQUE. Cette thèse a toujours fait l'objet de vives controverses et les adeptes de cette opinion passent aux yeux de certains comme des adversaires inconditionnels de tout risque. Je pense qu'un tel jugement est le résultat d'une notoire incompréhension de ce point de vue. En effet, l'emploi des stratégies de prudence ne contredit aucunement les stratégies de risque. Le risque est humain. Mais qu'est-il au juste ? C'est avant tout un système d'hypothèses sur une situation ou sur le comportement d'un autre joueur. En adoptant ce nouveau système décrivant une situation de risque, c'est-à-dire en rétrécissant d'une manière ou d'une autre l'ensemble des stratégies admissibles, on peut dans ces nouvelles conditions reformuler la stratégie minimax et estimer l'espérance.

Tout ce qui vient d'être dit concerne la partie « a-mathématique » de la recherche opérationnelle. Mais quelque importantes que soient l'intuition et la logique du « bon sens », l'analyse mathématique joue un rôle fondamental dans la conception du modèle et le choix du critère: détermination des estimations, de la structure de l'ensemble de Pareto et de nombreux autres problèmes auxiliaires de l'analyse de toute opération. Ceci est un dialogue systématique avec la nature, une chaîne d'expériences mathématiques. En analyse des systèmes et en recherche opérationnelle, ce sont précisément les expériences mathématiques qui remplacent les expériences physiques et les essais sur le modèle.

Les méthodes de résolution des problèmes d'optimisation occupent une place particulière parmi les méthodes mathématiques utilisées en analyse des systèmes. Elles constituent la clé de voûte du software des problèmes décisionnels. Ces méthodes qui relèvent de la programmation linéaire, non linéaire, dynamique et stochastique

ainsi que de l'optimisation discrète sont suffisamment bien élaborées. D'importants ouvrages leur sont consacrés et elles sont au programme des facultés préparant des spécialistes en mathématiques appliquées, en économie, en automatisation du projet, etc. Bien plus, des paquets de programmes réalisant diverses méthodes d'optimisation sont en voie de création ou ont déjà été créés dans tous les pays industrialisés. Ces paquets sont munis d'un programme de commande spécial qui permet de les utiliser en régime de dialogue et met ces méthodes d'optimisation à la portée des spécialistes qui ne sont pas des professionnels en mathématiques appliquées.

Les problèmes d'optimisation de la recherche opérationnelle possèdent généralement une caractéristique importante que l'analyste ne doit jamais perdre de vue. Nous avons déjà signalé que les critères doivent être stables pour les erreurs entraînées par la réalisation pratique des paramètres choisis. En d'autres termes, dans les problèmes d'optimisation correctement posés, il n'est pas nécessaire de chercher les valeurs optimales des paramètres avec une précision élevée.

Les raisonnements précédents nous conduisent à la conclusion très importante pour la réalisation pratique de la procédure d'analyse de l'opération. Une fois en possession d'un modèle détaillé et bien justifié, c'est-à-dire d'un système de contraintes et de liaisons entre ses éléments, écrit en langage mathématique, et une fois construit le critère $f(x)$, il ne faut pas se lancer immédiatement dans la résolution du problème d'extrémum

$$f(x) \Rightarrow \max_{x \in G}. \tag{5.1}$$

Ce problème peut être très difficile et nécessiter une grande perte de temps machine. Or, étant donné que pour choisir les paramètres il faut résoudre plusieurs fois un problème d'extrémum, le temps de résolution du problème (5.1) peut être un facteur décisif.

C'est pourquoi on aura intérêt à considérer de pair avec le modèle initial un modèle proche dans lequel les contraintes et le critère seront plus simples que dans (5.1). En d'autres termes, on propose de poser en même temps que (5.1) un problème « voisin »:

$$\varphi(x) \Rightarrow \max_{x \in H}. \tag{5.2}$$

Ce problème peut s'avérer bien plus simple que (5.1) et les valeurs optimales des indices du système seront pratiquement les mêmes. Désignons par $x^*$ la solution du problème (5.1) et par $\hat{x}$, celle du problème (5.2). Si le modèle simplifié est assez bon, alors la valeur $f(x^*)$ sera pratiquement la même que $f(\hat{x})$.

La construction des modèles simplifiés joue un grand rôle dans l'étude des opérations complexes. Pour choisir les paramètres, on

est généralement contraint de revenir à plusieurs reprises à des problèmes d'optimisation de type (5.1). Donc, pour mener à bien la recherche il est nécessaire que l'opération d'optimisation soit assez « rapide » et assez économique.

Il est donc question de deux classes de modèles et, partant, de deux classes d'algorithmes: les algorithmes rapides et les algorithmes vérificatifs. Les algorithmes rapides appliqués au modèle simplifié nous permettent de choisir les principaux paramètres de la future construction et de prendre les principales décisions relatives au projet. Les algorithmes rapides nous permettent en quelque sorte de faire une ébauche de l'opération: un avant-projet de la future construction ou un schéma général du futur projet. Ensuite on vérifie le déroulement de l'opération et les valeurs des caractéristiques de la construction à l'aide d'un modèle plus complet dont l'adéquation ne fait pas de doute. Si besoin est, on peut se servir de ce modèle pour procéder à une correction des décisions adoptées. Malheureusement il n'existe aucune procédure universelle de simplification du modèle. Quelques généralités seront développées à ce propos aux chapitres IV, V et VI. Mais dans la plupart des cas la simplification du modèle relève du bon sens.

Il est toujours difficile de justifier mathématiquement la possibilité de remplacer un modèle accepté par un modèle simplifié et d'utiliser les algorithmes rapides. Cette justification est parfois impossible notamment lorsque la simplification est liée à un rabaissement de l'ordre du système. Cependant le principe d'un processus décisionnel en deux étapes (et en plusieurs dans le cas de systèmes plus complexes), de la construction et de l'utilisation des algorithmes rapides et enfin de la combinaison des calculs effectués avec ces algorithmes et des calculs vérificatifs sur le modèle accepté est un principe majeur de l'analyse des systèmes. Ce principe apparaîtra en filigrane tout au long de l'ouvrage et certains chapitres même seront entièrement consacrés à ce problème et aux méthodes de construction d'algorithmes rapides.

Si l'analyse des systèmes tient ses principes méthodologiques et méthodiques de la recherche opérationnelle, ses outils d'investigation des problèmes dynamiques, elle les doit à la théorie de la commande, théorie qui fera l'objet du prochain chapitre.

# CHAPITRE II

# SYSTÈMES COMMANDÉS

## § 1. Remarques préliminaires

La théorie de la commande est au même titre que la recherche opérationnelle l'une des principales sources d'idées et de méthodes de l'analyse des systèmes moderne. C'est la première discipline scientifique entièrement centrée sur le développement des méthodes de prises de décision. La recherche opérationnelle et la théorie de la commande sont des disciplines très proches l'une de l'autre. Bien plus, la notion actuelle d'opération est tellement générale que tout problème de la théorie de la commande peut être posé en termes de recherche opérationnelle.

Formellement, la théorie de la commande, discipline qui couvre un vaste spectre de problèmes, peut être incluse dans la recherche opérationnelle. Mais cette généralisation n'est guère payante.

La théorie de la recherche opérationnelle étudie traditionnellement les problèmes de statique et à la limite les problèmes de décision en plusieurs étapes. La théorie de la commande, quant à elle, s'est concentrée dès le départ sur les problèmes de dynamique. Elle s'est de plus forgé un appareil si riche en idées originales et importantes pour les applications qu'il serait irrationnel d'identifier ces deux disciplines ou d'inclure l'une dans l'autre. Rétroaction, mouvement programmé, mécanisme de commande, commande optimale sont autant de notions élaborées par la théorie de la commande. Enfin, l'idée de la méthode de programmation, méthode qui devient aujourd'hui la base méthodologique de la gestion scientifique des processus sociaux, est, ainsi que nous le verrons plus bas, mise à profit en théorie de la commande. La décomposition du processus de commande en la commande du mouvement programmé suivie d'une correction par les mécanismes de commande qui est un élément important de la méthode de programmation et qui repose à la base de la synthèse de grands systèmes commandés est aussi une création de la théorie de la commande.

La théorie de la commande (ou théorie de la régulation, nom qu'elle portait les cent premières années après sa création) remonte aux années quarante du XIX$^{e}$ siècle, date à laquelle dans deux pays différents et indépendamment l'un de l'autre apparurent deux travaux consacrés au même problème: le choix des paramètres du régulateur de Watt. Les auteurs de ces travaux étaient le physicien anglais J.C. Maxwell et l'ingénieur russe I. Vychnégradski. Ces deux

travaux portaient sur un problème capital de l'époque, savoir l'élaboration des principes scientifiques de régulation du fonctionnement d'une machine à vapeur, plus précisément sur les principes permettant de stabiliser la vitesse de rotation de l'arbre qui est soumis à l'action d'une force extérieure variable. La nouvelle discipline s'est signalée par l'actualité des problèmes envisagés, une logique d'analyse irréprochable et la clarté des résultats acquis.

Au XIX$^e$ siècle, les problèmes de la théorie de la régulation mobilisèrent les plus éminents esprits scientifiques: A. Stodoll en Autriche-Hongrie, E. Routh en Angleterre, N. Joukovski en Russie. Les idées et méthodes de la théorie de la régulation furent fortement influencées par la théorie générale de la stabilité et principalement par la théorie de Liapounov dont le langage fut pendant longtemps le principal langage de la théorie de la commande.

Montrons comment la théorie de la régulation est rattachée à la théorie de la stabilité. Soit un processus physique commandé, c'est-à-dire un processus dont on peut changer le cours en faisant varier tel ou tel paramètre de construction. Considérons pour fixer les idées le projet de construction d'un pilote automatique, appareil qui est susceptible de modifier plusieurs fois la position des gouvernes, donc le caractère du vol de l'avion. Le pilote se voit fixer un but (par exemple le lieu et l'heure d'arrivée). L'itinéraire est défini en fonction de ce but. Le pilote place l'avion sur le cap et branche le pilote automatique. Le mouvement de l'avion sera décrit par un système d'équations différentielles

$$\dot{x} = f(x, t, p, \xi), \tag{1.1}$$

où $x$ est un vecteur décrivant l'état de phase du système, c'est-à-dire les coordonnées et la vitesse de l'avion, $\xi$, un vecteur aléatoire caractérisant les actions extérieures, $p$, le vecteur des paramètres de construction du pilote automatique, paramètres qui peuvent être fixés par le responsable de la réalisation de l'objectif par le pilote automatique (en l'occurrence le constructeur).

Lorsqu'on connaît la trajectoire de l'avion, supposée par convention stationnaire, on peut toujours choisir l'origine du temps de telle sorte qu'à cette trajectoire correspondent, en l'absence d'actions extérieures, les valeurs nulles des variables de phase. Donc le point $x = 0$ doit satisfaire l'équation

$$f(0, t, p, 0) \equiv 0. \tag{1.2}$$

Supposons par ailleurs qu'à un instant $t = t_0$, l'avoin soit attaqué par une perturbation aléatoire (par exemple un coup de vent) qui modifie l'état du système:

$$x(t_0) = x_0 \neq 0. \tag{1.3}$$

Quelles conditions doit remplir le mouvement de l'avion pour que ce dernier arrive au but fixé malgré l'écart (1.3)? Il est tout d'abord évident que les valeurs des composantes de la fonction vectorielle $x(t)$, caractérisant la position de l'avion par rapport à la trajectoire théorique, ne peuvent croître (en module). Etant donné que la trajectoire de l'avion doit passer par le but de la commande, il est nécessaire que l'écart provoqué par ces perturbations disparaisse à la longue. L'expérience montre qu'il suffit que le mouvement de l'avion soit asymptotiquement stable, c'est-à-dire que

$$\lim_{t\to\infty} x(t) = 0. \tag{1.4}$$

REMARQUE. La condition (1.4) n'est ni strictement nécessaire ni strictement suffisante. En effet, le but de la commande doit être atteint au bout d'un intervalle de temps $T$ fini, alors que la condition (1.4) décrit les propriétés asymptotiques du mouvement pour $t \to \infty$. La possibilité de remplacer un horizon fini par un infini ne signifie qu'une chose: le temps de vol d'un avion est « pratiquement » infini et il est bien plus grand que le temps nécessaire à la compensation de la perturbation (temps d'amortissement des oscillations de l'avion). Même si la possibilité d'utiliser la condition (1.4) est un fait empirique, on verra plus bas qu'elle ouvre des voies diverses à l'analyse quantitative. Le changement d'un horizon grand mais fini par un infini est couramment utilisé, car l'étude des fonctions est plus aisée pour $t$ infini que pour $t$ fini.

L'étude de la stabilité de la solution triviale ($x = 0$) du système (1.1) constitue le problème fondamental de la théorie de la stabilité, l'un des plus importants chapitres de la théorie des équations différentielles. Mais la construction d'un pilote automatique assurant à l'avion un mouvement asymptotiquement stable ne se ramène pas uniquement à un problème de théorie de la stabilité. D'une façon générale, il ne suffit pas d'établir la stabilité du vol d'un avion avec un pilote automatique donné. Bien plus, il faut généralement déterminer les intervalles admissibles de variations des paramètres du pilote automatique (les coordonnées du vecteur $p$) responsables de la stabilité, c'est-à-dire résoudre un problème, dans un certain sens, inverse.

Vu que les écarts par rapport à la trajectoire théorique — les quantités $x^i(t)$ — doivent être petits, on résoudra ce problème par linéarisation, c'est-à-dire qu'on remplacera l'équation (1.1) par une équation linéaire. En négligeant les termes d'ordre $O(x^2)$, on obtient

$$\dot{x} = Ax, \tag{1.5}$$

où $A$ est la matrice $\left(\frac{\partial f}{\partial x}\right)_{x=0,\,\xi=0}$. L'équation (1.5) est homogène (en vertu de la condition (1.2)) et l'étude des propriétés des solutions de l'équation (1.5) est considérablement simplifiée.

Au stade initial de son développement, la théorie de la régulation étudiait les problèmes de régulation de mouvements stationnaires: fonctionnement d'une machine à un régime constant, mouvement

rectiligne uniforme d'un avion à une altitude constante donnée, etc. Dans le cas de mouvements stationnaires, le second membre de l'équation (1.1) ne contient pas la variable temps, et les éléments de la matrice $A$ sont constants. Dans ces conditions, le problème de la stabilité asymptotique se ramène à un pur problème d'algèbre: trouver les conditions que doivent satisfaire les coefficients de l'équation caractéristique $|A - \lambda E| = 0$ pour que toutes ses racines soient à parties réelles strictement négatives. Ce problème s'appelle *problème de Routh-Hurwitz* ; son étude a jeté les fondements de la théorie de la régulation automatique *).

Les conditions nécessaires de stabilité asymptotique de la solution triviale de l'équation (1.5) peuvent être établies de plusieurs manières. Par exemple, multiplions scalairement ses deux membres par le vecteur $x$; l'équation devient alors

$$\frac{d}{dt}(x, x) = 2(x, Ax), \tag{1.6}$$

où (. , .) désigne le produit scalaire.

Si la forme quadratique

$$(x, Ax) = \sum_{i, j} a_{ij} x^i x^j,$$

où $a_{ij} = \left.\frac{\partial f^i}{\partial x^j}\right|_{x=0,\ \xi=0}$ sont les éléments de la matrice $A$, est définie positive, c'est-à-dire $(x, Ax) > 0$ pour tout $x \neq 0$, alors le produit scalaire $(x, x)$ croît et la solution triviale ne peut être asymptotiquement stable. Aussi pour conditions de stabilité utilise-t-on souvent les conditions de définition négative de la forme quadratique $(x, Ax)$. Aux termes du critère de Sylvestre, ces conditions peuvent être mises sous la forme

$$a_{11} < 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \ \ldots, \ (-1)^n \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix} > 0. \tag{1.7}$$

Etant donné que les valeurs $a_{ij}$ dépendent des paramètres de construction du régulateur, c'est-à-dire des coordonnées du vecteur $p$, les conditions (1.7) peuvent être traitées comme les conditions régissant le choix de $p$.

Les conditions de stabilité peuvent être représentées sous des formes différentes et le choix de l'une d'elles est subjectif. Des questions de commodité et de suggestivité ont poussé les ingénieurs

*) Les conditions de Routh-Hurwitz sont exposées dans n'importe quel ouvrage de théorie de la commande automatique (cf. par exemple [10]).

à chercher des formes toujours nouvelles de représentation des conditions de stabilité: critères de Mikhaïlov, de Nyquist, etc.

Les conditions de stabilité écrites dans une forme quelconque définissent un ensemble $G_1$ dans l'espace des paramètres. Le choix de $p \in G_1$ assure la réalisation de la condition de stabilité (1.4) et partant la réalisation du but de la commande. Mais la condition de stabilité n'est pas la seule condition que doivent remplir les paramètres du régulateur. L'établissement de cette condition repose toujours sur un schéma concret, sur une certaine structure. Ainsi les travaux de Maxwell et de Vychnégradski étudiaient la stabilité du régime d'un moteur à l'aide d'une structure bien concrète: le régulateur de Watt. Donc, lorsqu'on étudie les procédés de commande d'un processus, on doit toujours résoudre simultanément deux problèmes: le choix d'un régulateur (d'un pilote automatique ou d'un autre mécanisme) et ensuite le choix des paramètres réalisant le but de la commande. Le premier problème fait toujours appel à l'ingéniosité. Bien qu'il existe de nombreuses recherches d'où sont sorties des recommandations diverses sur le choix de la structure des mécanismes de commande, il n'empêche qu'en fin de compte le problème du schéma du mécanisme est un problème de construction et l'appréciation de la perfection de cette construction ne se ramène pas à la seule caractéristique qui nous intéresse présentement. Simplicité, technicité, sûreté sont des facteurs aussi importants. Donc, le schéma de construction impose aux paramètres $p$ des contraintes bien définies:

$$p \in G_2, \tag{1.8}$$

et en outre la condition (1.8) est le plus souvent de la forme

$$p_-^i \leqslant p^i \leqslant p_+^i, \tag{1.9}$$

où $p^i$ est la $i$-ième coordonnée du vecteur $p$. Donc, pour réaliser le but de la commande, c'est-à-dire les conditions (1.4), il est nécessaire que

$$p \in G = G_1 \cap G_2. \tag{1.10}$$

Il est possible que l'ensemble $G$ soit vide. Cela voudra dire que le régulateur choisi (le système de commande) n'est pas en mesure de réaliser le but de la commande et doit être remplacé par un autre.

Signalons que le problème du choix des valeurs des paramètres de construction qui vient d'être décrit est, de par sa position, proche des problèmes de recherche opérationnelle étudiés au chapitre précédent et peut, en outre, être formulé en termes de recherche opérationnelle. En effet, dans ce problème on a un but de commande qui peut être représenté sous forme de la condition (1.4), on a un système de contraintes (1.1) et enfin les ressources indispensables à la réalisation du but, c'est-à-dire la possibilité de choisir les paramètres du régulateur satisfaisant des contraintes de la forme (1.8).

Le critère de réalisation du but de la commande (1.4) peut également être formulé en termes d'optimisation. Il suffit à cet effet d'introduire un nouveau critère, par exemple:

$$J(p)=\begin{cases}1 & \text{si } \lim\limits_{t\to\infty} x(t)=0,\\ 0 & \text{si } \lim\limits_{t\to\infty} x(t)\neq 0,\end{cases} \qquad (1.11)$$

ou sous la forme équivalente

$$J(p)=\begin{cases}1 & \text{si } p\in G,\\ 0 & \text{si } p\notin G.\end{cases} \qquad (1.11')$$

La condition (1.8) ou (1.10) peut maintenant être écrite sous la forme suivante:

$$J(p) \Rightarrow \max, \qquad (1.12)$$

et toute solution du problème d'optimisation (1.12) est solution du problème initial. La reformulation du problème en termes d'optimisation ne modifie visiblement pas sa nature. Elle fait ressortir d'avantage les liens génétiques des problèmes de la recherche opérationnelle et de ceux de la théorie de la commande.

Les conditions (1.4) et (1.12) ne définissent pas une seule solution du problème. En fait, elles déterminent tout un ensemble de paramètres $G$ (toute une classe de constructions admissibles) qui réalisent le but de la commande. Donc, le constructeur a encore la possibilité de préciser le vecteur $p$, de le soumettre à des contraintes supplémentaires, ce qui équivaut à une optimisation supplémentaire, mais sur l'ensemble $G$ et non plus sur $G_2$.

Dès les années trente furent publiés les premiers travaux consacrés à l'étude de la qualité de la commande, c'est-à-dire au choix dans un ensemble de commandes stables d'une commande qui satisfasse des conditions supplémentaires. Dans le cas du pilotage automatique d'un avion de ligne, cette condition subsidiaire pourrait être par exemple la minimisation des surcharges: le passager ne se contente pas uniquement d'un vol stable qui l'amène à destination, il lui faut aussi un certain confort. S'il existe plusieurs moyens d'arriver au but, alors il est possible d'offrir le confort désiré. Le problème de l'appréciation de la qualité de la régulation peut donc être formulé dans les termes suivants: sur un ensemble de paramètres $p$ vérifiant la condition

$$p \in G$$

trouver les valeurs de $p$ qui réalisent la condition

$$I(p) \Rightarrow \min, \qquad (1.13)$$

où $I(p)$ est une fonctionnelle dépendant de la trajectoire et caractérisant par exemple le niveau des surcharges tolérables. D'une façon

générale, l'appréciation de la qualité de la régulation n'est pas un problème simple. Désignons par $w$ le vecteur des accélérations, par $T$ le temps de vol de l'avion. Il y a lieu alors de considérer quelques critères susceptibles chacun d'estimer d'une manière ou d'une autre la surcharge, c'est-à-dire le confort du passager, par exemple:

$$I_1(p) = \max_{t \in [0, T]} |w|, \tag{1.14}$$

$$I_2(p) = \int_0^T |w| \, dt, \tag{1.15}$$

$$I_3(p) = \int_0^T (w, w) \, dt. \tag{1.16}$$

Le confort du passager et les surcharges sont liés par une relation qui est loin d'être univoque et il est peu probable qu'on puisse la caractériser par un seul nombre. Le critère (1.14) est une caractéristique locale et de toute évidence il est souhaitable que le système de commande « coupe » toutes les surcharges en crête au cours du vol. Il est non moins souhaitable de diminuer les estimations intégrales des surcharges, décrites par les critères (1.15) ou (1.16). L'estimation de la qualité de la commande est donc un problème type à plusieurs critères et en l'étudiant on se heurte au même cortège d'indéterminations que dans le chapitre précédent.

Dans les revues techniques des années d'avant-guerre furent publiés d'innombrables articles consacrés à la discussion des critères de qualité d'une commande. On étudiait le plus souvent des problèmes avec un critère (1.16). Mais je pense que ce choix était dicté par le fait que les critères (1.14) et (1.15) conduisaient à des problèmes si compliqués qu'il n'y avait aucun espoir de les résoudre analytiquement. On constate donc que dès les années trente la théorie de la régulation était confrontée à des problèmes qui par leurs positions étaient parfaitement identiques aux problèmes actuels de la recherche opérationnelle bien que les auteurs qui les étudiaient étaient loin de partager les regards actuels quant à leur contenu.

Dans l'après-guerre on assiste à un élargissement qualitatif du spectre des problèmes de la théorie de la régulation, élargissement qui a relégué au second plan les problèmes traditionnels. A la fin de ce processus, le terme de « théorie de la régulation » disparaît et est remplacé par celui de « théorie de la commande ».

Un tel élargissement du cercle des problèmes étudiés était lié principalement aux besoins des toutes nouvelles techniques spatiales. Avant la guerre, la théorie de la régulation s'intéressait en principe aux processus qui s'étendaient sur un long intervalle de temps. C'est précisément pour cette raison que les méthodes de la théorie de

la stabilité classique qui étudie les propriétés asymptotiques des solutions (pour $t \to \infty$) se prêtent bien à l'étude des problèmes réels: la durée du vol d'un avion est bien plus élevée que le temps de compensation de ses mouvements d'oscillation. La théorie de la régulation a même introduit la notion de processus transitoires ou processus de retour d'un système au régime stationnaire initial à la fin de l'action d'une perturbation aléatoire.

La théorie de la régulation qui avait essentiellement pour objet les régimes stationnaires ne s'intéressait aux processus transitoires que dans le cadre de l'étude de la qualité de la commande. Les besoins des techniques spatiales ont conduit à des problèmes fondamentalement différents, puisque le mouvement des fusées est en principe de courte durée et peut être traité comme un processus transitoire unique. Mais il y avait encore une circonstance qui impliquait la position de nouveaux problèmes: c'était le coût du carburant. Le rapport de la quantité de carburant consommé pour le transport d'une cargaison au poids de cette cargaison est généralement si élevé que le calcul de la trajectoire la plus économique devint dès milieu des années 40 l'un des plus importants problèmes de la théorie mathématique du pilotage des missiles. Ce problème a intéressé un grand nombre de mathématiciens et d'ingénieurs dont les efforts ont donné naissance à une nouvelle discipline scientifique, la théorie de la commande optimale (cf. [6, 8, 9]). Parmi les travaux consacrés à ce sujet, on retiendra surtout celui de D. Okhotsimski [56] qui, à la terminologie près, posait les problèmes actuels de la théorie de la commande optimale.

Les problèmes traités par la théorie des missiles diffèrent considérablement des problèmes traditionnels de la théorie de la régulation. Nous avons déjà signalé que dans les années quarante, les efforts des spécialistes furent concentrés essentiellement sur l'étude des procédés de commande de mouvements stationnaires sur un intervalle de temps infini. Les problèmes de la dynamique des missiles sont, quant à eux, essentiellement non stationnaires. De plus, le processus de commande est souvent de très courte durée. Par exemple, le temps de combustion des charges de poudre des premières fusées de combat se chiffrait en secondes ou en dixièmes de seconde et le processus de commande se déroulait durant cet intervalle de temps. Les positions traditionnelles des problèmes de la théorie de la commande ne pouvaient en aucun cas correspondre aux nouvelles réalités. Il fallait une refonte des méthodes de raisonnement des experts, leur recyclage vers des problèmes fondamentalement nouveaux. Pour toutes ces raisons la dynamique du pilotage des missiles se développa pendant assez longtemps en marge de la théorie de la régulation. La fusion avec la théorie de la régulation qui devait donner naissance à la théorie de la commande eut lieu plus tard, dans les années cinquante.

5—01523

En théorie de la régulation, le problème de la recherche du mouvement programmé ne se posait généralement pas. Ce mouvement soit était connu à l'avance (par exemple le nombre de tours d'une turbine à la seconde), soit sa détermination était triviale (par exemple le calcul des paramètres de vol d'un avion à une altitude donnée). La théorie de la régulation, on l'a vu, étudiait un autre problème : la construction de l'opérateur de rétroaction assurant le régime donné (un mouvement programmé donné). Ces deux problèmes — la détermination du mouvement programmé et sa commande — sont certes étroitement liés (et ce lien sera discuté dans le détail). Mais jusqu'à un certain temps les bases théoriques de résolution de ces problèmes se développèrent indépendamment l'une de l'autre. Le calcul des trajectoires programmées se transforma en une importante branche de la théorie de la commande en gestation. Ces problèmes relevaient généralement du calcul des variations. Le travail de D. Okhotsimski déjà signalé inaugura cette voie. Plus tard, en U.R.S.S. d'abord et aux U.S.A. ensuite, furent publiées les premières recherches fondamentales consacrées à ces problèmes. Les travaux de T. Enéev et L. Chatrovski en U.R.S.S., de G. Leitmann aux U.S.A. et de nombreux autres jetèrent les bases de cette nouvelle branche.

Au carrefour des années quarante et cinquante, la théorie de la régulation classique commence à s'intéresser aux problèmes variationnels et principalement aux problèmes de rapidité. La création de cette nouvelle branche en théorie de la régulation est généralement rattachée au nom de A. Feldbaum qui publia de nombreux articles sur les problèmes de rapidité au début des années cinquante.

Les problèmes variationnels qui se posèrent à la dynamique des missiles et à la théorie de la régulation possédaient une particularité importante qui ne permettait pas de les ramener à des problèmes variationnels classiques et de leur appliquer ensuite les méthodes élaborées de l'analyse mathématique. On sait que les problèmes classiques du calcul des variations se formulent aisément en termes de théorie de la commande. La difficulté était ailleurs. Le calcul des variations opérait — pour employer le langage de la théorie de la commande — sur des commandes $u(t)$ qui n'étaient soumises à aucune contrainte ou dont les ensembles admissibles étaient ouverts. Dans la dynamique des missiles et en théorie de la régulation, un problème fondamentalement nouveau se posait : les commandes $u(t)$ devaient en principe appartenir à des ensembles fermés, traduisant ainsi des besoins techniques naturels. La poussée du moteur est en principe bornée :

$$p(t) \leqslant p_{\max} \ \forall t. \tag{1.17}$$

De même, l'angle de rotation du gouvernail et les intervalles de variation des autres organes de commande satisfont toujours des

conditions de type (1.17). Cette particularité exclut la possibilité d'utiliser directement les méthodes bien élaborées du calcul des variations et impliqua la création d'un appareil spécial. Même si au début des années cinquante de nombreux problèmes de ce type furent résolus, aucune méthode générale d'analyse ne put être dégagée. Pour chaque problème, on développait une méthode en propre. Néanmoins on peut rapporter au début de ces années cinquante la formation de la théorie connue aujourd'hui sous le nom de théorie de la commande optimale. Cette théorie était appelée à servir de base à l'actuelle théorie de la commande de systèmes techniques dont les idées et méthodes ont fortement marqué l'orientation de l'analyse des systèmes.

L. Pontriaguine joua un rôle éminent dans le développement de la théorie de la commande optimale en formulant le principe du maximum qui permet grâce aux multiplicateurs de Lagrange de ramener un problème de commande optimale à un problème aux limites spécial pour un système d'équations différentielles ordinaires (cf. [9]).

Les travaux de L. Pontriaguine et de son école en théorie de la commande consacrèrent le langage et les méthodes et modelèrent l'actuelle théorie de la commande. La communauté du langage et du formalisme contribua à l'unification de la théorie de la régulation et des autres branches qui étudiaient les problèmes de commande.

Aujourd'hui le principe de l'optimisation et les méthodes de la théorie de la commande optimale sont largement utilisés dans les recherches appliquées, les projets et exercent une influence décisive sur le développement de l'analyse des systèmes.

## § 2. Notion de systèmes commandés

Au paragraphe précédent on a donné un bref historique des idées fondamentales de la théorie de la commande. On se propose maintenant de concrétiser quelques notions déjà utilisées et de leur conférer le sens qui nous sera indispensable pour la suite de l'exposé. Les systèmes commandés seront traités comme le développement naturel des systèmes dynamiques dont les propriétés sont supposées connues à partir de la théorie des équations différentielles.

a) *Commandes*. On se bornera à l'étude de systèmes décrits par des équations différentielles ordinaires

$$\dot{x} = f(x, u, t, \xi), \tag{2.1}$$

où $x$ est un vecteur de phase de dimension $n$, $\xi$, le vecteur des perturbations (actions extérieures) de dimension $k$ ($k \leqslant n$) qui peut être aléatoire (auquel cas il est défini par ses caractéristiques probabilistes) ou indéfini (exprimant ainsi l'insuffisance de nos connaissances

sur le phénomène étudié). Dans les deux cas le vecteur $\xi(t)$ est défini par son appartenance à un certain ensemble:

$$\xi(t) \in G_\xi(t) \quad \forall t. \tag{2.2}$$

La fonction vectorielle $u(t)$ de dimension $m \leqslant n$ s'appelle *commande* ou *vecteur commande*. On admet qu'au système est associé un responsable capable et habilité à prendre des décisions, c'est-à-dire à choisir la commande qui peut être fonction du temps ($u = u(t)$), du vecteur de phase ($u = u(x)$), des perturbations ($u = u(\xi)$) ou être de forme plus générale ($u = u(t, x, \xi)$).

En théorie de la commande qui met aussi en jeu des systèmes techniques le terme « responsable » désignera aussi le constructeur qui conçoit le système de commande.

Dans tous les cas où la commande est une fonction des variables de phase et des perturbations, on admet que ces quantités sont connues ou seront connues du responsable au moment de la prise de décision. On verra plus bas que cette hypothèse implique à son tour la description d'un processus informationnel. La commande $u$ est généralement soumise à des contraintes que nous écrirons sous la forme

$$u \in G_u \qquad \forall t, x, \xi, \tag{2.3}$$

où $G_u$ est un ensemble arbitraire.

Les coordonnées de phase peuvent également être assujetties à des contraintes, par exemple:

$$x \in G_x \qquad \forall t. \tag{2.4}$$

Les conditions (2.3) et (2.4) sont parfois regroupées:

$$(t, x, u) \in G_{xu} \qquad \forall t \tag{2.5}$$

ou

$$(t, x, u, \xi) \in G \qquad \forall t. \tag{2.6}$$

La condition (2.4) s'appelle *contrainte de phase*, les conditions (2.5) et (2.6), *contraintes mixtes*. Les conditions mixtes s'appellent souvent contraintes de type *goulot d'étranglement*. Ce terme a été emprunté à l'économie; par exemple, la production d'un bien ne peut excéder la capacité d'une usine, capacité qui à son tour dépend des facteurs de gestion, c'est-à-dire de l'affectation d'une partie des biens aux investissements. Les recherches effectuées en mathématiques de l'économie exercent une grande influence sur le développement de la théorie de la commande moderne.

On appellera *systèmes commandés* des systèmes d'équations de la forme (2.1) avec les contraintes (ensembles) $G_\xi$, $G_u$, $G_x$. La signification de cette notion sera examinée ultérieurement dans le cadre d'exemples concrets.

REMARQUE. En plus des systèmes (2.1) on a parfois à considérer leurs analogues aux différences, par exemple

$$x_{n+1} = x_n + \tau f(x_n, u_n, \xi_n, t_n), \tag{2.1'}$$

ainsi que les systèmes de forme plus générale

$$x_{n+1} = F_n(x_n, u_n, \xi_n). \tag{2.1''}$$

Les équations discrétisées de la forme (2.1′) ou (2.1″) ne se présentent pas uniquement lors de la discrétisation du problème, qui est nécessaire à son analyse sur ordinateur. Il existe des systèmes à temps discret qu'il est difficile de ramener à la forme (2.1). Tel est le cas par exemple des systèmes décrivant la dynamique des populations à générations non croisées, des processus agricoles, etc. Donc, les systèmes de la forme (2.1), que l'on étudiera essentiellement, ne sont pas les seuls systèmes importants pour la pratique.

b) *But de la commande et critère de qualité.* L'assertion suivante est répandue parmi les spécialistes en théorie de la commande: « Il n'y a pas de commande sans objectif ». Nous verrons plus loin que cette assertion n'est pas absolue, de même d'ailleurs que la plupart des assertions de cette nature liées à des problèmes concrets, mais on peut provisoirement la poser pour axiome. De sorte que par commande on sous-entendra le processus de formation des objectifs, la quête et la mise en œuvre des méthodes de leur réalisation.

Les systèmes commandés sont conçus pour réaliser tel ou tel objectif: l'avion, pour le transport de passagers ou de marchandises, la fusée, pour mettre un engin spatial sur une orbite donnée; un système de gestion économique, pour fournir des biens à la population, etc. L'objectif de la commande est une notion subjective de la personne responsable du choix des commandes sur les motivations qui la guident dans le choix de la fonction $u(\cdot)$.

Une fois choisie, la fonction $u(t, x, \xi)$ est une description formalisée des procédés de réalisation de l'objectif. En théorie de la commande cette fonction est souvent appelée *loi de la commande.* Donc, un problème fondamental de la théorie de la commande est la quête de la loi de la commande en fonction de l'objectif donné.

On sait que l'objectif de la commande peut être formulé en termes de maximisation d'une fonctionnelle:

$$J_1(u) \Rightarrow \max. \tag{2.7}$$

Supposons par exemple qu'il s'agisse de placer un engin spatial sur une orbite circulaire à un instant donné $t = T$:

$$\| x \|_{t=T} = R, \tag{2.8}$$

où $\| x \| = \sqrt{(x^1)^2 + (x^2)^2 + (x^3)^2}.$

La fonctionnelle (2.7) peut être donnée de différentes manières. Par exemple

$$J_1(u)=\begin{cases}1 & \text{si } (x^1)^2+(x^2)^2+(x^3)^2\,|_{t=T}=R,\\ 0 & \text{si } (x^1)^2+(x^2)^2+(x^3)^2\,|_{t=T}\neq R.\end{cases} \tag{2.7'}$$

Mais la maximisation de la fonction (2.7′) n'est pas l'unique objectif du responsable chargé du lancement de la fusée. Son aspiration légitime est de réaliser cet objectif avec des dépenses minimales. Cet objectif peut être formalisé en termes de maximum d'une fonctionnelle de la forme

$$J_0(u)=\int_0^T F(x,\ u)\,dt. \tag{2.9}$$

On remarquera que toutes les contraintes de la forme (2.2) à (2.6) peuvent être aussi énoncées dans les mêmes termes. En effet, définissons la fonctionnelle $J_2(u)$, par exemple, de la manière suivante:

$$J_2(u)=\begin{cases}1 & \text{si } (t,\ x,\ u,\ \xi)\in G,\\ 0 & \text{si } (t,\ x,\ u,\ \xi)\notin G.\end{cases}$$

La condition (2.6) peut alors être mise sous la forme

$$J_2(u)\Rightarrow \max. \tag{2.10}$$

Signalons enfin que l'équation (2.1) est elle-même une contrainte, car l'exigence que l'évolution du système, c'est-à-dire la variation du vecteur $x$, soit régie par l'équation (2.1) peut aussi être formulée en termes d'optimisation.

Donc, le processus de commande avec des objectifs fixes se ramène à la détermination d'une commande $u(\cdot)$ qui réalise simultanément le maximum d'un ensemble de fonctionnelles

$$J_i(u)\Rightarrow \max, \qquad i=0,\ 1,\ 2,\ \ldots,\ k. \tag{2.11}$$

Ce problème n'admet une signification mathématique que dans certains cas spéciaux. Désignons par $\Omega_i$ l'ensemble des commandes qui réalisent le maximum des fonctionnelles $J_i$, c'est-à-dire l'ensemble des fonctions vectorielles qui sont solutions des problèmes $J_i(u)\Rightarrow \max$. Pour que le problème d'optimisation (2.11) ait un sens, il est alors nécessaire que

$$\bigcap_{i=0}^{k}\Omega_i\neq\varnothing, \tag{2.12}$$

c'est-à dire que l'intersection des ensembles de commandes réalisant les maximums des fonctionnelles ne soit pas vide.

Si la condition (2.12) n'est pas remplie, le problème de commande ne se ramène pas directement à un problème de maximisation de

type (2.11) et dans ce cas il est nécessaire de passer par une analyse non formelle du problème initial (comme en recherche opérationnelle).

En théorie de la commande il existe plusieurs moyens pour tourner cette difficulté. Tout d'abord on définit l'objectif de la commande et on fixe les contraintes. Bien que formellement, ainsi que nous l'avons déjà vu, toutes les conditions qui doivent être satisfaites par la solution cherchée peuvent être écrites sous une forme unique à l'aide de l'opération de maximisation, il est plus commode de distinguer l'objectif de la commande (la fonctionnelle $J_1$), les contraintes (les fonctionnelles $J_2, \ldots, J_k$) et la fonctionnelle de qualité (la fonctionnelle $J_0$). Ces conditions possèdent des significations physiques différentes et leur séparation apporte la clarté qui permet à l'ingénieur ou à l'analyste de poser le problème dans la forme la mieux appropriée aux objectifs fixés à telle ou telle construction.

En théorie de la commande des systèmes techniques, un cas classique est celui où

$$\omega = \bigcap_{i=1}^{k} \Omega_i \neq \varnothing,$$

c'est-à-dire celui où l'on peut réaliser l'objectif de la commande en satisfaisant toutes les contraintes. C'est précisément ce cas qui a été étudié par la théorie de la régulation. L'ensemble $\omega$ n'étant pas vide, il est naturel d'y chercher une solution qui maximise encore une fonctionnelle: la fonctionnelle de qualité $J_0$.

Donc, le problème (2.11) est généralement remplacé par le suivant:

$$J_0(u) \Rightarrow \max_{u \in \omega} \tag{2.13}$$

c) *Exemple.* Considérons un exemple qui a assez fortement marqué les idées et méthodes de la théorie moderne de la commande: le problème de mise sur orbite d'un engin spatial. Dans la suite de l'exposé nous reviendrons souvent à cet exemple pour clarifier les diverses particularités de la théorie.

L'équation du mouvement du centre de masse d'une fusée est de la forme

$$m\frac{d^2x}{dt^2} = mg + R + Q, \tag{2.14}$$

où $x$ est un vecteur de phase, $m = m(t)$, la masse de l'engin, $g$, le vecteur intensité des forces de pesanteur, $R = R(x, dx/dt)$, la force aérodynamique agissant sur l'appareil, $Q = qku$, la force de réaction, $q$ (scalaire), la perte de masse:

$$\frac{dm}{dt} = q, \tag{2.15}$$

$u$, le vecteur unitaire de la direction de la force de réaction, $k$, un coefficient dépendant des paramètres de construction du moteur et avant tout de la vitesse d'écoulement des gaz. Les quantités $q$ et $u$ sont des commandes qui doivent en outre satisfaire les contraintes

$$0 \leqslant q \leqslant q_{\max}, \tag{2.16}$$

$$\| u \|^2 = (u^1)^2 + (u^2)^2 + (u^3)^2 = 1. \tag{2.17}$$

L'objectif de la commande se formule comme suit: l'engin doit être placé sur une orbite circulaire à un instant $T$, c'est-à-dire que les variables de phase doivent satisfaire les conditions

$$x(T) = x_T, \qquad \left.\frac{dx}{dt}\right|_{t=T} = V_T. \tag{2.18}$$

La première de ces conditions définit le point orbital en lequel doit se trouver l'engin, la deuxième signifie qu'au moment où il atteint ce point sa vitesse doit être telle qu'après l'extinction des moteurs il tourne sur une orbite circulaire donnée.

Le vecteur de phase $x(t)$ est soumis à la contrainte évidente:

$$\|x(t)\| \geqslant R_g \qquad \forall t > 0, \tag{2.19}$$

où $R_g$ est le rayon du globe terrestre.

De plus, la mise sur orbite doit être réalisée avec une consommation minimale de carburant ou, ce qui revient au même ici, la plus grande masse doit être placée sur orbite, autrement dit

$$m(T) = m_0 - \int_0^T q\,dt \Rightarrow \max. \tag{2.20}$$

Ainsi, le problème de la mise sur orbite d'un engin spatial se ramène à la recherche de commandes $q(t)$ et $u(t)$ vérifiant les conditions (2.14) à (2.19) et réalisant le maximum de la fonctionnelle (2.20). La commande qui vérifie toutes les conditions $J_i \Rightarrow \max$, pour $i = 1, 2, \ldots, k$, c'est-à-dire réalise l'objectif de la commande et satisfait les contraintes, sera dite *commande admissible*. Dans l'exemple envisagé, si le temps $T$ est assez élevé, il existe une infinité de commandes admissibles qui constituent l'ensemble $\omega$ sur lequel est considéré le problème (2.13).

Le problème de la mise sur orbite d'un engin spatial qui vient d'être formulé n'avait pas d'analogue en théorie classique de la régulation. Sa résolution donnait une trajectoire d'appui, ou comme il est d'usage de dire maintenant, une trajectoire programmée. On peut considérer que l'analogue du mouvement commandé en théorie de la régulation est le régime stationnaire que doit maintenir le mécanisme de commande (le régulateur de Watt, le pilote automatique, etc.).

## § 3. Problème stochastique et optimisation en deux étapes

L'exemple que nous avons traité dans le paragraphe précédent est dans un certain sens exceptionnel, car il se rapporte à une situation déterministe pure qui n'existe pratiquement pas dans la réalité.

Si la fonction vectorielle $\xi = \xi(t)$ du second membre de l'équation (2.1) est aléatoire, alors il en est de même du vecteur de phase $x = x(t)$ aussi. Donc, le problème général de théorie de la commande est un problème de commande d'un processus aléatoire (stochastique). La position générale de ce problème est excessivement compliquée. Des méthodes variées de simplification de ce problème ont été mises au point en théorie de la commande. D'une façon générale, on tente de ramener l'analyse des systèmes stochastiques réels à l'étude successive d'une série de problèmes déterministes. Certes, cette réduction est loin d'être universelle (voire même pas toujours possible) mais elle est très importante en théorie de la commande (et en analyse des systèmes). Dans ce paragraphe, on cite un exemple type de réduction d'un problème stochastique à un problème déterministe.

a) *Sur la position du problème.* Soit donc un système commandé stochastique assez général

$$\dot{x} = f(t, x, u, \xi). \tag{3.1}$$

Supposons que notre objectif est de faire passer le système pendant un intervalle de temps $T$ de l'état $x(0) = x_0$ à l'état $x(T) = x_T$. Le système (3.1) étant soumis à l'action de forces aléatoires $\xi(t)$, quelle que soit la commande, le vecteur de phase $x(t)$ sera aussi une fonction aléatoire du temps. Nous ne pouvons donc pas énoncer l'objectif de la commande en termes déterministes, et la condition $x(T) = x_T$ doit être remplacée par une autre, formulée en termes probabilistes, par exemple

$$J_1 = \overline{(x(T) - x_T)^2} \Rightarrow \min \tag{3.2}$$

ou

$$J_2 = P\{\|x(t) - x_T\| < \varepsilon\} \Rightarrow \max, \tag{3.3}$$

où $P\{y < a\}$ est la probabilité que la quantité aléatoire $y$ soit strictement inférieure à la quantité déterministe $a$ *).

La condition (3.2) minimise la variance pour $\overline{x(T)} = x_T$ et (3.3) maximise la probabilité qu'une valeur finie du vecteur de phase se trouve à l'intérieur d'un intervalle assez petit de centre $x_T$. Ces

*) On rappelle que la barre désigne l'espérance mathématique de la quantité qu'elle surmonte.

deux conditions estiment d'une manière ou d'une autre la précision avec laquelle est réalisé l'objectif fixé ($x(T) = x_T$).

REMARQUE. Ces représentations probabilistes de l'objectif de la commande n'épuisent aucunement les éventuelles interprétations de la précision avec laquelle est réalisé l'objectif de la commande. Par exemple, on peut remplacer la fonctionnelle (3.2) par la fonctionnelle

$$J_3 = \overline{((x(T) - x_T), \quad R(x(T) - x_T))}, \tag{3.2'}$$

où $R$ est une matrice définie positive dont les coefficients introduisent une relation d'équivalence des estimations de la précision de réalisation des conditions $x^i(T) = x_{iT}$, $i = 1, \ldots, n$.

Outre les contraintes probabilistes de la forme (3.2), (3.3), dans le problème peuvent intervenir des contraintes déterministes *), par exemple des contraintes sur les variables de phase

$$x \in G_x,$$

ou sur la commande

$$u \in G_u$$

ou

$$\int_0^T u\,dt \leqslant C, \tag{3.4}$$

etc.

Donc, en décrivant les problèmes stochastiques de la théorie de la commande, on est confronté à la même situation d'existence des contraintes déterministes « physiques » que dans la recherche opérationnelle.

La fonctionnelle de qualité de la commande

$$J_0 = \int_0^T F(x, u)\,dt \tag{3.5}$$

se formule aussi en général sous la forme déterministe.

Les expressions de la forme (3.4) sont correctes: la quantité de carburant dans une fusée est toujours bornée par une quantité bien définie. Mais l'estimation de la qualité de la commande sous la forme (3.5) n'est déjà plus suffisamment justifiée: il est évident que cette qualité peut être estimée par une caractéristique moyenne, par exemple par la quantité

$$J_0^* = \int_0^T \overline{F(x, u)}\,dt. \tag{3.6}$$

*) Par contraintes déterministes on entendra ici et dans la suite des conrai ntes qui sont presque sûrement réalisées.

Donc, un problème de commande présente de nombreuses difficultés spécifiques qui ne permettent de le ramener à un problème d'optimisation que par l'introduction d'hypothèses supplémentaires. La première difficulté posée par cette réduction est due au fait que le problème contient des contraintes stochastiques et des contraintes déterministes. Pour ces problèmes il n'existe pas de théorie mathématique assez générale et à plus forte raison d'algorithmes de calcul. La deuxième, qui est plus importante encore, réside dans le fait qu'un problème stochastique est par essence à plusieurs critères: il ne se réduit pas à un problème de la forme (2.13). Donc, en analysant un problème stochastique, nous devons nécessairement utiliser des hypothèses.

La théorie de la commande étudie les problèmes stochastiques dans des positions variées reflétant les diverses particularités des systèmes envisagés, qui permettent de formuler les hypothèses simplificatrices.

b) *Schéma d'optimisation en deux étapes*. Les situations dans lesquelles les perturbations sont supposées petites donnent lieu à une importante classe de problèmes de commande. L'hypothèse que $\xi \equiv 0$ doit alors donner une première approximation satisfaisante pour la résolution du problème.

Sous cette condition la fonction $x(t)$ ne sera plus un processus aléatoire et les fonctionnelles (3.2), (3.2′) se transforment en expressions finies. Par exemple, la fonctionnelle (3.2) devient

$$J_1 = \overline{(x(T) - x_T)^2} = (x(T) - x_T)^2.$$

Cette fonctionnelle atteint son minimum pour

$$x(T) = x_T. \tag{3.7}$$

On est donc conduit à un ordinaire problème de commande optimale qui consiste à déterminer une commande $u(t)$ vérifiant des contraintes déterministes et faisant passer le système

$$\dot{x} = f(x, u, t, 0)$$

de l'état $x(0) = x_0$ à l'état $x(T) = x_T$ de manière à maximiser une fonctionnelle de qualité:

$$J_0 = \int_0^T F(x, u)\, dt \Rightarrow \max. \tag{3.8}$$

La trajectoire $x(t)$, solution de ce problème, s'appelle *trajectoire programmée* (ou *optimale*) et la commande $u(t)$ qui la réalise, *commande programmée* (ou *optimale*).

Revenons à l'exemple de la mise sur orbite d'un engin spatial. La situation dans laquelle nous avons décrit ce problème dans le

paragraphe précédent nous donne précisément une trajectoire programmée. Ce problème nous permet de déterminer la quantité de carburant nécessaire à la mise sur orbite et de nous faire une idée du caractère de la trajectoire de l'engin spatial. Nous pouvons aussi trouver les lois de commande: les fonctions $q(t)$ et $u(t)$ décrivant, la première, le caractère de consommation du carburant, la seconde, les variations en fonction du temps de l'inclinaison des déflecteurs de jet qui commandent la direction de la poussée de réaction.

Mais si l'on tente de se servir de la loi de commande optimale trouvée pour résoudre le problème posé, on ne pourra pas atteindre l'état final donné à la date fixée. La trajectoire programmée est une trajectoire qui n'est jamais suivie par la fusée. En effet, le mouvement de la fusée (et de tout autre système commandé) est défini non seulement par les commandes, mais aussi par l'action de facteurs aléatoires incontrôlables. Dans le cas du mouvement d'une fusée par exemple, ces facteurs sont les fluctuations de la densité de l'air, les coups de vent, les erreurs d'exécution des commandes, etc. Leur action entraîne une déviation de la fusée de la trajectoire programmée. Il est donc nécessaire d'envisager un dispositif dont la mission est de neutraliser les effets aléatoires. Convenons d'appeler ce système de correction pilote automatique. Il est évident que les principes de construction de ce système de correction des mécanismes de commande diffèrent fondamentalement de ceux qui régissent la construction d'une commande programmée. Du point de vue mathématique ce sera un problème tout autre. Entrons dans le vif du sujet.

Soit donc à considérer de nouveau le système (3.1). Si l'on suppose que $\xi \equiv 0$, on obtient alors

$$\dot{x} = f(t, x, u, 0). \tag{3.9}$$

En résolvant le problème (3.7), (3.8) pour ce système, on trouve la trajectoire et la commande programmées

$$x = \hat{x}(t), \qquad u = \hat{u}(t). \tag{3.10}$$

La trajectoire et la commande réelles différeront, en vertu de l'hypothèse de petitesse des perturbations aléatoires, peu de leurs valeurs programmées. Pour cette raison, on posera

$$x = \hat{x}(t) + y(t), \qquad u = \hat{u}(t) + v(t). \tag{3.11}$$

Linéarisons le système (3.1) en y portant (3.11) sous l'hypothèse que $y$ et $v$ sont du même ordre de petitesse que $\xi$. On obtient ainsi le système linéaire

$$\dot{y} = Ay + Bv + C\xi, \tag{3.12}$$

où $A = \left(\frac{\partial f}{\partial x}\right)_0$, $B = \left(\frac{\partial f}{\partial u}\right)_0$, $C = \left(\frac{\partial f}{\partial \xi}\right)_0$ sont des matrices; les dérivées sont calculées ici le long de la trajectoire programmée, c'est-à-dire pour les valeurs nulles de $\xi$, $v$ et $y$. Sans nuire à la généralité on peut admettre que le processus aléatoire $\xi(t)$ est centré, c'est-à-dire que

$$\overline{\xi(t)} = 0 \qquad \forall t. \tag{3.13}$$

Résolvons le système (3.12) par la méthode de variation des constantes arbitraires. Désignons par

$$Y = (y_i^j)$$

la matrice des solutions particulières linéairement indépendantes du système d'équations homogène

$$\dot{y} = Ay, \tag{3.14}$$

système en vertu duquel la matrice $Y$ est solution de l'équation matricielle

$$\frac{dY}{dt} = AY, \quad Y(0) = E. \tag{3.15}$$

où $E$ est la matrice unité.

Cherchons la solution de l'équation (3.12) sous la forme

$$y = Yc,$$

où $c$ est un vecteur dont les coordonnées sont des fonctions du temps inconnues. En portant cette expression dans (3.12) et en se servant de (3.15), on trouve que le vecteur $c$ est solution de l'équation différentielle

$$Y\dot{c} = Bv + C\xi,$$

ou

$$c = \int_0^t Y^{-1}(\tau)(Bv + C\xi)\,d\tau + c^*,$$

où $c^*$ est le vecteur des constantes d'intégration arbitraires.

L'inconnue $y$ devant satisfaire des conditions initiales nulles, on a $c^* = 0$ et l'on trouve

$$y = \int_0^t G(t, \tau)(Bv + C\xi)\,d\tau, \tag{3.16}$$

où $G$ est la matrice de Green ($G(t, \tau) = Y(t)\,Y^{-1}(\tau)$).

Pour trouver une nouvelle commande, on se sert de la condition (3.2) que l'on peut désormais mettre sous la forme

$$\overline{(y(T),\ y(T))} \Rightarrow \min. \tag{3.17}$$

Grâce à (3.16), on peut écrire (3.17) sous la forme

$$J_1=\overline{(y(T),\ y(T))}=\int_0^T\int_0^T G(t,\ s)\,G(t,\ \tau)\,\{\overline{(Bv(s),\ Bv(\tau))}+$$
$$+\overline{(Bv(s),\ C\xi(\tau))}+\overline{(Bv(\tau),\ C\xi(s))}+\overline{(C\xi(s),\ C\xi(\tau))}\}\,d\tau. \quad (3.18)$$

Supposons maintenant que l'on cherche à déterminer la commande correctrice sous la même forme $v = v(t)$ que la commande programmée, c'est-à-dire déterminer une fonction déterministe du temps qui minimiserait la variance (3.17). La fonction $v(t)$ ne sera plus aléatoire et par suite

$$\overline{(Bv(\tau),\ Bv(s))}=(Bv(\tau),\ Bv(s)),$$
$$\overline{(Bv(s),\ C\xi(\tau))}=(Bv(s),\ \overline{C\xi(\tau)})=0,$$
$$\overline{(Bv(\tau),\ C\xi(s))}=(Bv(\tau),\ \overline{C\xi(s)})=0,$$

l'expression (3.18) devient alors:

$$J_1=\overline{(y(T),\ y(T))}=\int_0^T\int_0^T (B(s)\,v(s),\ B(\tau)\,v(\tau))\,d\tau\,ds+$$
$$+\int_0^T\int_0^T \overline{(C(s)\,\xi(s),\ C(\tau)\,\xi(\tau))}\,d\tau\,ds. \quad (3.19)$$

Chaque terme du second membre de (3.19) est strictement positif. Si donc l'on veut choisir une commande $v(t)$ de façon à minimiser (3.19), il nous faut nécessairement prendre $v \equiv 0$, autrement dit, si l'on veut réduire l'erreur à l'aide d'une fonction du temps donnée, le meilleur moyen de commander le système est de ne pas le commander du tout.

Il faut donc choisir la commande correctrice $v$ telle qu'elle dépende d'une manière ou d'une autre des perturbations $\xi(t)$. Mais il est difficile de choisir directement $v$ comme une fonction de $\xi$, car on doit pouvoir mesurer cette quantité. Mais on peut indirectement tenir compte de la dépendance $v(\xi)$ en mesurant les variables de phase ou leurs écarts par rapport au mouvement programmé, c'est-à-dire qu'on doit chercher la commande correctrice sous la forme

$$v = v(t,\ x) \qquad (\text{ou } v = v(t,\ y)). \quad (3.20)$$

Donc, la recherche de la commande correctrice optimale se ramène à celle d'une fonction $v = v(t,\ x)$ minimisant la fonctionnelle (3.17) sous la condition (3.12). De plus, la commande $v(t,\ x)$ est soumise aux contraintes suivantes:

$$u = (\hat{u} + v) \in G_u.$$

Ce problème est différent de celui de commande programmée et s'appelle problème de synthèse ou problème de construction de l'opérateur de rétroaction. Si dans le problème de commande programmée la commande cherchée dépend seulement de l'état initial du système et de l'objectif, dans le problème de synthèse, en revanche, elle doit être définie pour toutes les valeurs du vecteur de phase $x(t)$.

La détermination de l'opérateur de rétroaction était un problème fondamental de la théorie de régulation. Mais pour le trouver il fallait utiliser la condition de stabilité. Dans la commande stochastique, la construction de l'opérateur de rétroaction conduit à des problèmes d'optimisation spéciaux dont un cas particulier est de toute évidence le problème de stabilité.

c) *Justification du schéma d'optimisation en deux étapes.* Ainsi l'hypothèse de petitesse des perturbations extérieures nous permet non seulement de décomposer le problème en deux sous-problèmes moins compliqués, mais aussi de répartir nos possibilités de commande de manière à résoudre deux problèmes d'optimisation. Dans la première étape nous avons choisi la commande programmée, laquelle nous a permis de déterminer la trajectoire programmée qui assure la réalisation de l'objectif donné (en l'absence de perturbations extérieures) avec des dépenses minimales. Dans la deuxième étape, nous avons défini les paramètres du pilote automatique (la fonction $v(t, x)$), dispositif qui avec les ressources disponibles assure la plus grande précision possible de réalisation de l'objectif. Ce schéma de résolution du problème de commande s'appelle souvent *optimisation en deux étapes*. Il joue un rôle important non seulement en théorie de la commande de systèmes techniques, mais, comme nous le verrons plus bas, il est à la base de la méthode de programmation de systèmes économiques et autres dont le fonctionnement est déterminé par les activités des gens.

Le schéma développé est de toute évidence une méthode euristique de résolution du problème à deux critères, car en principe il n'existe pas de procédure régulière rigoureusement justifiée pour le résoudre. Mais puisque ce schéma nous donne rapidement une solution bien définie, il est nécessaire de l'estimer en la comparant aux solutions obtenues par d'autres méthodes de choix de la loi de commande $v(t, x)$.

La procédure exposée nous donne une règle pour le choix de la solution dans un problème à deux critères

$$J_0(u) \Rightarrow \min, \qquad J_1(u) \Rightarrow \min.$$

Supposons que nous avons composé une convolution linéaire de ces fonctionnelles de la forme

$$I(u) = J_0(\cdot) + cJ_1(\cdot), \tag{3.21}$$

où $c > 0$ est une constante fixe.

Considérons maintenant le problème

$$I(u) \Rightarrow \min.$$

Montrons que l'on peut toujours encadrer la fonctionnelle $I(u)$. Désignons par $\hat{u}$ et $v$ les solutions des deux sous-problèmes mentionnés dans ce paragraphe. La fonction $u^* = \hat{u} + v$ étant une solution admissible, on a

$$I^* \leqslant I(\hat{u} + v), \tag{3.22}$$

où $I^* = I(u^*)$ désigne la valeur minimale de la fonctionnelle (3.21) réalisée par la fonction $u = u^*$.

D'autre part, il est immédiat de prouver que toujours

$$I^* \geqslant J_0(\hat{u}). \tag{3.23}$$

En effet, supposons par absurde que

$$J_0(\hat{u}) > I^*, \tag{3.24}$$

et calculons

$$I^* = I(u^*) = J_0(u^*) + cJ_1(u^*).$$

Les fonctionnelles $J_0$ et $J_1$ étant strictement positives de par leur signification, il est évident que $I^* \geqslant J_0(u^*)$. En vertu de (3.24), on a alors $J_0(\hat{u}) > J_0(u^*)$. Mais cette inégalité contredit l'hypothèse que $\hat{u}$ est une commande optimale minimisant la fonctionnelle $J_0(\cdot)$. On obtient ainsi l'encadrement

$$J_0(\hat{u}) \leqslant I^* \leqslant J_0(\hat{u} + v) + cJ_1(\hat{u} + v). \tag{3.25}$$

L'encadrement (3.25) montre que la méthode de choix de la solution fondée sur une optimisation en deux étapes fournit une précision satisfaisante pour toutes les convolutions de la forme (3.21), si seulement la valeur de la fonctionnelle $J_1(\cdot)$ n'est pas trop élevée ou si la constante $c$, qui caractérise la « qualité » de la fonctionnelle $J_1(\cdot)$ pour le responsable, ne l'est pas non plus. Le schéma d'optimisation en deux étapes qui sont le choix du programme et la construction du mécanisme de réalisation de ce programme est l'un des plus importants procédés euristiques de la théorie de la commande moderne. Il est dans une égale mesure nécessaire à la commande de systèmes techniques et technologiques et à la gestion de systèmes sociaux et économiques où il est devenu la principale méthode de programmation des commandes. L'encadrement (3.25) justifie cette méthode. Plus exactement, il permet d'étudier les possibilités d'application de l'optimisation en deux étapes à l'analyse de systèmes concrets, c'est-à-dire de remplacer la résolution du problème donné (l'optimisation de la fonctionnelle (3.21)) par celle de deux sous-problèmes d'optimisation.

Les prochains paragraphes seront consacrés à la réalisation numérique du schéma d'optimisation en deux étapes.

## § 4. Méthodes de calcul des programmes optimaux utilisant le principe du maximum

Un programme optimal se définit comme la solution du problème variationnel

$$J(u) \Rightarrow \min \tag{4.1}$$

avec les conditions

$$\dot{x} = f(x, u, t), \tag{4.2}$$

$$(x, u) \in G. \tag{4.3}$$

Le problème (4.1), (4.2), (4.3) s'appelle *problème de commande optimale* (cf. par exemple [6, 9]).

Les méthodes de résolution des problèmes d'extrémum de fonctions se répartissent généralement en deux classes: les méthodes directes et les méthodes utilisant les conditions nécessaires. Bien que conventionnelle, cette classification est néanmoins commode, car elle s'appuie sur la différence des approches dans la recherche de l'extrémum. Illustrons ceci sur l'exemple élémentaire de détermination du minimum libre d'une fonction $g(x)$.

Les méthodes directes se fondent sur l'idée de la pente (par exemple du gradient):

$$x_{n+1} = x_n - \alpha_n \frac{\partial g}{\partial x}(x_n),$$

quant à la deuxième classe de méthodes, elle est liée à la possibilité de réduction du problème d'extrémum à la résolution d'une équation transcendante

$$\frac{\partial g}{\partial x} = 0$$

Chacune de ces approches a ses vertus et ses défauts et son propre domaine d'application.

Les méthodes de calcul des programmes optimaux sont aussi réparties en deux classes. La première est composée des méthodes directes qui utilisent une réduction du problème variationnel (4.1), (4.2), (4.3) à un problème finidimensionnel. Cette discrétisation donne lieu à un problème spécial de programmation mathématique. Ces méthodes ont connu une diffusion particulièrement intense ces derniers temps. Elles sont à la base d'innombrables programmes standards et de paquets de programmes à usage divers et en particulier des systèmes d'optimisation dialoguants (cf. [8]). Ces méthodes feront l'objet du prochain paragraphe.

Cependant les méthodes directes ne sont pas universelles et il existe de nombreux problèmes pour lesquels elles sont peu efficaces. Expliquons-nous sur l'exemple d'un schéma élémentaire de réduc-

tion d'un problème de commande optimale à un problème de programmation mathématique. Supposons que la fonctionnelle $J(u)$ est de la forme

$$J(u)=\int_0^T F(x,\ u,\ t)\,dt. \qquad (4.4)$$

Subdivisons l'intervalle d'intégration en $N$ intervalles partiels de longueur $\tau$ sur lesquels nous admettrons que la commande est constante. Remplaçons l'équation (4.2) par l'élémentaire schéma aux différences (schéma d'Euler)

$$x_{k+1} = x_k + \tau f(x_k,\ u_k,\ t_k). \qquad (4.5)$$

Faisons de même avec la fonctionnelle (4.4):

$$J(u)=\tau\sum_{k=0}^{N-1} F(x_k,\ u_k,\ t_k). \qquad (4.6)$$

Particularisons la condition (4.3). Supposons, par exemple, que

$$x(0) = x_0,\ x(T) = x_N,\ u_k \in G_k,\ k = 0,\ \ldots,\ N, \qquad (4.7)$$

où $G_k$ sont des ensembles convexes.

La minimisation de la fonctionnelle (4.6) sous les contraintes (4.5) et (4.7) est un problème classique de programmation non linéaire. Les difficultés de résolution de ce problème tiennent à d'innombrables facteurs dont les plus importants sont sa dimension, le nombre de variables et le nombre de contraintes. La dimension du problème dépend essentiellement de deux circonstances: de la dimension du vecteur $x$ et du nombre d'intervalles partiels $N$, c'est-à-dire du produit $n \times N$.

Si $N$ est assez grand, alors même dans le cas où la dimension du vecteur $x$ n'est pas trop élevée, le problème de programmation mathématique est excessivement compliqué. Le nombre $N$ est défini avant tout par des impératifs de précision. Dans les problèmes techniques, il est très fréquent que $N$ soit très grand (dans les problèmes de dynamique du vol, de mécanique du milieu continu, etc.). Dans de telles situations, il est nécessaire de faire appel à d'autres méthodes peu sensibles à l'accroissement du nombre $N$ d'intervalles partiels. Telles sont les méthodes utilisant les conditions nécessaires d'extrémum, sauf que contrairement à la recherche de l'extrémum libre d'une fonction à l'aide des conditions nécessaires, le problème pour la fonctionnelle se ramène non pas à la recherche des zéros de la fonction $\partial g(x)/\partial x$, mais à un problème aux limites pour des équations différentielles ordinaires.

Le principe du maximum joue un rôle particulier en théorie des méthodes numériques de calcul des programmes optimaux. Formu-

lons ce principe et définissons ses possibilités sur l'exemple d'une fonctionnelle de la forme (4.4). Considérons le système (4.2) et son adjoint

$$\dot{\psi} = -\frac{\partial H}{\partial x} = \varphi(x, \psi, u, t), \tag{4.8}$$

où

$$H = (\psi, f) - F(x, u, t) \tag{4.9}$$

est le hamiltonien.

Supposons que la commande est soumise à la contrainte

$$u(t) \in G \quad \forall t. \tag{4.10}$$

On a alors le

THÉORÈME (principe du maximum de Pontriaguine). *Pour qu'une fonction $u(t)$ minimise la fonctionnelle* (4.4) *sous les conditions* (4.2) *et* (4.10), *il est nécessaire et suffisant qu'elle maximise le hamiltonien, c'est-à-dire qu'elle soit solution du problème*

$$H(x, \psi, u, t) \Rightarrow \max_{u \in G}. \tag{4.11}$$

Il faut ajouter des conditions de transversalité à la condition (4.11). Nous envisagerons essentiellement des problèmes dans lesquels nous fixerons l'état initial du système:

$$x(0) = x_0. \tag{4.12}$$

Considérons d'abord le cas où l'état final du système est aussi fixé:

$$x(T) = x_T. \tag{4.13}$$

Dans cette situation, le théorème formulé permet de ramener directement le problème de recherche d'un programme optimal à un problème aux limites pour un système d'équations différentielles ordinaires. En effet, on peut à partir de la condition (4.11) déterminer la commande $u$ comme une fonction de $\psi$, $x$ et $t$:

$$u = u(\psi, x, t). \tag{4.14}$$

En portant l'expression (4.14) dans les équations (4.2) et (4.8) on obtient le système suivant d'équations d'ordre $2n$ ($n$ est la dimension du vecteur $x$):

$$\dot{x} = X(x, \psi, t), \quad \dot{\psi} = \Psi(x, \psi, t), \tag{4.15}$$

où $X(x, \psi, t) = f(x, u(x, \psi, t), t)$, $\Psi(x, \psi, t) = \varphi(x, \psi, u(x, \psi, t), t)$.

Pour le système (4.15) nous avons exactement $2n$ conditions aux limites: $n$ conditions scalaires (4.12) à l'extrémité gauche de la trajectoire et $n$ conditions (4.13) à l'extrémité droite.

Comme il est question de conditions nécessaires, le principe du maximum de Pontriaguine dit que le programme optimal ne peut se trouver que parmi les solutions du problème aux limites (4.12), (4.13) pour le système (4.15).

Si l'extrémité droite de la trajectoire n'est pas fixée et qu'elle soit soumise à des conditions en nombre inférieur à $n$, alors pour ramener le problème à un problème aux limites il faut encore poser un certain nombre de conditions, appelées *conditions de transversalité*. Ces conditions portent sur les valeurs aux bornes des impulsions (des variables adjointes) $\psi(T)$.

Elucidons la signification des conditions de transversalité sur quelques exemples.

Supposons par exemple que l'extrémité droite de la trajectoire n'est soumise à aucune contrainte. Dans ce cas les variables adjointes doivent satisfaire la condition

$$\psi(T) = 0. \tag{4.16}$$

La condition (4.16) « ferme » le problème: nous avons de nouveau $2n$ conditions: $n$ conditions (4.12) à l'extrémité gauche et $n$ conditions (4.16) à l'extrémité droite.

La condition (4.16) est un cas particulier des conditions de transversalité. On peut poser des conditions analogues; si par exemple la trajectoire doit finir sur la surface

$$\Phi(x(T)) = 0, \tag{4.17}$$

alors pour $t = T$ la fonction $\psi(t)$ doit satisfaire une condition de transversalité de la forme

$$\psi(T) = \lambda \frac{\partial \Phi}{\partial x}, \tag{4.18}$$

où $\lambda$ est un facteur scalaire qui est choisi à partir de la condition (4.17). Nous sommes de nouveau conduits à un problème aux limites pour le système (4.15) avec $n$ conditions sur l'extrémité gauche et $n$ sur l'extrémité droite. La déduction détaillée des conditions de transversalité est accessible dans tout ouvrage de calcul variationnel ou de théorie de commande optimale.

Ainsi, grâce aux conditions nécessaires la recherche du programme optimal se ramène à un problème aux limites pour un système d'équations différentielles ordinaires. La procédure de résolution de ce problème contient à titre d'étape intermédiaire la résolution d'un problème d'extrémum auxiliaire (4.11) consistant à maximiser le hamiltonien. Ce problème, nous devons le résoudre à chaque pas de l'intégration numérique et sa dimension $m$ est celle du vecteur commande, autrement dit, le théorème de Pontriaguine décompose un problème de dimension $m \times N$ en $N$ problèmes de dimension $m$ reliés par une procédure d'intégration numérique d'équations diffé-

rentielles. Au fur et à mesure que $N$ croît, cette approche prend le pas sur les méthodes directes, car les difficultés de résolution des problèmes d'optimisation de la programmation non linéaire croissent exponentiellement, alors que celles des problèmes aux limites ne croissent que linéairement avec $N$. Par ailleurs, ces méthodes sont sensibles à l'accroissement de la dimension du vecteur $x$. Donc, les méthodes directes de calcul des programmes optimaux et les méthodes basées sur la réduction, grâce au principe du maximum, d'un problème de commande optimale à un problème aux limites pour un système d'équations différentielles ordinaires, se complètent mutuellement.

Le seul problème de théorie des équations différentielles pour lequel il existe des procédures numériques de résolution bien élaborées est le problème de Cauchy (méthodes d'Euler, de Runge-Kutta, etc.). Ces procédures font défaut aux problèmes aux limites et les diverses méthodes mises au point pour les résoudre utilisent d'une manière ou d'une autre la réduction à une suite de problèmes de Cauchy.

Les méthodes de résolution des problèmes aux limites de la théorie de la commande optimale peuvent être conventionnellement réparties en trois classes.

a) *Méthode d'ajustage*. Cette méthode est parfois appelée méthode de sélection des conditions initiales.

Une particularité des problèmes aux limites envisagés dans ce paragraphe est qu'à l'extrémité gauche de la trajectoire les valeurs des coordonnées de phase sont données tandis que les valeurs des impulsions $\psi(0)$ sont inconnues. Aussi la première chose qui vient à l'esprit est de tenter de choisir les valeurs initiales des impulsions $\psi(0)$ de sorte à satisfaire les conditions à l'extrémité droite. Pour fixer les idées ces conditions seront les conditions (4.13).

Supposons que nous ayons donné $\psi(0) = \psi_0$ par un procédé quelconque. Nous connaissons désormais les données initiales du système (4.2), (4.8) et nous pouvons résoudre le problème de Cauchy par une méthode quelconque. Servons-nous par exemple de la méthode d'Euler. Donnons-nous le pas d'intégration $\tau$ et posons

$$x((k+1)\tau) = x(k\tau) + \tau f(x(k\tau), u(k\tau), k\tau), \qquad (4.19)$$
$$\psi((k+1)\tau) = \psi(k\tau) + \tau\varphi(x(k\tau), u(k\tau), k\tau).$$

En particulier,

$$\begin{aligned} x(\tau) &= x(0) + \tau f(x(0), u(0), 0), \\ \psi(\tau) &= \psi(0) + \tau\varphi(x(0), u(0), 0). \end{aligned} \qquad (4.20)$$

Pour déterminer $x(\tau)$ et $\psi(\tau)$ il nous faut trouver la commande $u(0)$. A cet effet, en vertu du principe du maximum, il faut trouver

$$\max_{u(0)\in G_0} H(\psi(0), f(x(0), u(0), 0)). \qquad (4.11')$$

Si $G_0$ est un ouvert, alors pour résoudre le problème (4.11′) on peut se servir des conditions nécessaires de maximum, conditions qui nous permettent de ramener ce problème à la résolution de l'équation transcendante

$$\frac{\partial H}{\partial u}=0. \tag{4.21}$$

Après avoir déterminé $u$ (0) à partir de la condition (4.11′), on peut calculer $x$ ($\tau$) et $\psi$ ($\tau$) à l'aide des formules (4.20). En reprenant cette procédure on trouve successivement les valeurs $x$ ($2\tau$), $\psi$ ($2\tau$); $x$ ($3\tau$), $\psi$ ($3\tau$), etc. On obtient finalement les coordonnées $x^i$ ($T$) du vecteur $x$ ($T$). Les quantités $\psi$ (0) ayant été arbitrairement choisies, les conditions $x^i(T)-x^i_T=0$, $i=1, \ldots, n$, ne seront pas généralement satisfaites pour $t=T$. Introduisons le vecteur d'écart

$$\Phi = x(T) - x_T.$$

Il est évident que $\Phi$ sera fonction de $\psi$ (0): $\Phi=\Phi(\psi(0))$.

Le problème sera résolu si l'on choisit un vecteur $\psi$ (0) tel que

$$\Phi(\psi(0))=0. \tag{4.22}$$

Notre méthode nous ramène donc à un problème classique de détermination des zéros d'une fonction vectorielle. Pour le résoudre on peut se servir de méthodes bien élaborées (par exemple, la méthode de Newton) pour lesquelles il existe d'innombrables paquets de programmes appliqués. Il semblerait donc qu'il n'y ait aucune complication particulière du moins lorsque le vecteur $x$ est de dimension peu élevée. En fait, la situation est bien plus complexe et la réalisation de la procédure décrite peut parfois se heurter à d'insurmontables difficultés. La dépendance fonctionnelle $\Phi$ ($\psi$ (0)) se réalise par la résolution du problème de Cauchy. Or ce problème est en raison des particularités du système (4.2), (4.8) toujours instable. (J'insiste bien sur le mot toujours.) Illustrons ceci sur l'exemple simple où l'équation (4.2) est linéaire en $x$:

$$\dot{x}=Ax+f_1(u). \tag{4.23}$$

On distinguera deux cas: 1) la partie réelle de l'une au moins des valeurs propres de la matrice $A$ est strictement positive; 2) les parties réelles de toutes les valeurs propres de la matrice $A$ sont strictement négatives.

Dans le premier cas, les composantes de la solution de l'équation (4.23) seront à croissance exponentielle quelle que soit la fonction $u$ ($t$). Cela signifie que la résolution numérique du problème de Cauchy nous met dans l'obligation d'opérer avec des nombres élevés et la solution sera instable.

Dans le deuxième cas, on se heurte aux mêmes difficultés lors de l'intégration de l'équation adjointe (4.8). Cette équation sera de la

forme

$$\dot{\psi} = -\frac{\partial H}{\partial x} = -A^*\psi + \ldots, \tag{4.24}$$

où $A^*$ désigne la matrice transposée. Les valeurs propres des matrices $A$ et $A^*$ étant les mêmes, à chaque racine à partie réelle $< 0$ de l'équation $|A - \lambda E| = 0$ correspondra une racine à partie réelle $> 0$ de l'équation $|-A^* - \lambda E| = 0$. Si donc la solution de l'équation (4.23) est stable, celle de l'équation (4.24) sera instable et inversement, c'est-à-dire que cette procédure nous met inévitablement devant des difficultés de calcul dues à l'instabilité de la solution du problème de Cauchy. Le calcul de la fonction $\Phi(\psi(0))$ sera de plus compliqué.

L'instabilité de la solution du problème de Cauchy fait qu'un faible écart des valeurs initiales $\psi(0)$ par rapport aux valeurs qui auraient dû figurer dans la solution finale conduit à des valeurs très élevées de $\Phi(\psi(0))$. Pour cette raison la réalisation de toute procédure numérique de la méthode d'ajustage exige préalablement une analyse poussée qui permette de choisir une bonne approximation initiale. Par « bonne » on entend une approximation initiale pour laquelle la valeur de l'écart $\Phi(\psi(0))$ n'est pas trop grande. En se servant par exemple des modifications de la méthode de Newton, on peut conduire les calculs avec une grande précision.

Malgré ses inconvénients la méthode d'ajustage est très largement répandue et elle a permis de résoudre un grand nombre de problèmes techniques et économiques.

b) *Méthode de réduction à un problème linéaire.* Il existe une classe de problèmes de calcul des programmes optimaux pour lesquels peuvent être proposées des méthodes de résolution efficaces. Dans ces problèmes, la recherche du programme optimal se ramène à la résolution d'un problème de Cauchy sans procédure supplémentaire de détermination des zéros d'une fonction à l'exemple du numéro précédent. C'est une classe de problèmes linéaires de commande optimale à fonctionnelle quadratique.

Considérons un système régi par l'équation linéaire

$$\dot{x} = Ax + Bu, \tag{4.25}$$

où $A$ et $B$ sont des matrices dont les éléments $a_{ij}$ et $b_{ij}$ sont des fonctions données du temps. Sous la forme scalaire, le système (4.25) s'écrit :

$$\dot{x}^i = \sum_{j=1}^{n} a_{ij}x^j + \sum_{j=1}^{m} b_{ij}u^j. \tag{4.25'}$$

Supposons que l'état initial est donné :

$$x^i(0) = x_0^i, \qquad i = 1, \ldots, n \tag{4.26}$$

et cherchons la commande $u(t)$ qui fait passer le système de l'état (4.26) pendant l'intervalle de temps $T$ à l'état

$$x^i(T) = x_T^i, \qquad i = 1, \ldots, n, \tag{4.27}$$

et qui réalise le minimum de la fonctionnelle

$$J = \int_0^T [(x, Cx) + (x, Du) + (u, Eu)]\, dt \tag{4.28}$$

sur la trajectoire $x(t)$. Ici $C = (c_{ij})$, $D = (d_{ij})$, $E = (e_{ij})$ sont des matrices dont l'ordre est visiblement défini par les dimensions des vecteurs $x$ et $u$; $E$ est une matrice définie positive.

Composons le hamiltonien

$$H = (\psi, Ax + Bu) - (x, Cx) - (x, Du) - (u, Eu).$$

L'équation des impulsions sera de la forme:

$$\dot{\psi} = -\frac{\partial H}{\partial x} = -A^*\psi + \hat{C}x + Du, \tag{4.29}$$

où $\hat{C} = C + C^*$.

La commande n'étant soumise à aucune contrainte, pour que le hamiltonien atteigne son maximum sur la trajectoire optimale, il est nécessaire que

$$\frac{\partial H}{\partial u} = B^*\psi - D^*x - \hat{E}u = 0, \tag{4.30}$$

où $\hat{E} = E + E^*$. Si la matrice $E$ n'est pas singulière, c'est-à-dire si $E^{-1}$ existe, alors le système d'équations (4.30) admet une solution unique:

$$u = \hat{E}^{-1}(B^*\psi - D^*x). \tag{4.31}$$

Dans ce cas donc nous obtenons la commande sous la forme explicite d'une fonction des impulsions et des variables de phase et en outre cette fonction est linéaire.

En portant l'expression (4.31) dans le système d'équations (4.25), (4.29), on ramène ce dernier à la forme

$$\dot{x} = M_1 x + N_1 \psi, \qquad \dot{\psi} = M_2 x + N_2 \psi, \tag{4.32}$$

où $M_1 = A - B\hat{E}^{-1}D^*$, $N_1 = B\hat{E}^{-1}B^*$, $M_2 = \hat{C} - D\hat{E}^{-1}D^*$, $N_2 = -A^* + D\hat{E}^{-1}B^*$. Donc, dans le cas le plus général d'une fonctionnelle quadratique le problème du calcul d'un programme optimal pour un système linéaire se ramène à un problème aux limites pour un système d'équations différentielles linéaires.

Les équations différentielles linéaires constituent la seule classe d'équations différentielles pour lesquelles sont élaborées des méthodes de résolution des problèmes aux limites. Expliquons leur con-

tenu. Considérons l'équation

$$\dot{y} = Ay + f, \tag{4.33}$$

où $y$, $f$ sont des fonctions vectorielles de dimension $n$ et soit à résoudre l'équation (4.33) sous les conditions

$$(l_i,\ y(t_i)) = \sum_j l_i^j y^j(t_i) = \alpha_i, \quad i = 0,\ 1,\ \ldots,\ k. \tag{4.34}$$

La méthode qui permet de résoudre efficacement le problème (4.33), (4.34) est basée sur la possibilité de transférer les conditions aux limites (4.34) d'un point à un autre. On dira que la condition (4.34) est transférée d'un point $t_i$ en un point arbitraire $t$ si l'on arrive à déterminer (indépendamment de $y$) une fonction vectorielle $g(t)$ et une fonction scalaire $\beta(t)$ vérifiant les conditions

$$g(t_i) = l_i, \quad \beta(t_i) = \alpha_i, \tag{4.35}$$

de telle sorte que $(g(t), y(t)) = \beta(t)$ à tout instant $t \neq t_i$. Il est immédiat de s'assurer que pour cela on peut se servir de l'équation adjointe. Par équation adjointe on entendra l'équation

$$\dot{g} = -A^* g, \tag{4.36}$$

où $A^*$ est la transposée de la matrice $A$.

Multiplions scalairement les deux membres de l'équation (4.33) par $g$ et l'équation (4.36), par $y$, et ajoutons les résultats obtenus:

$$(\dot{y},\ g) + (y,\ \dot{g}) = (Ay,\ g) + (g,\ f) - (A^*g,\ y). \tag{4.37}$$

Comme $(Ay, g) = (A^*g, y)$, l'équation (4.37) peut être mise sous la forme

$$d(y,\ g)/dt = (g,\ f),$$

d'où

$$(g,\ y)_t = (g,\ y)_{t_i} + \int_{t_i}^{t} (g,\ f)\, dt.$$

Cette égalité prouve l'important théorème suivant.

THÉORÈME. *Si une fonction* $g(t)$ *satisfait l'équation* (4.36) *et la première des conditions* (4.35) *et une fonction* $\beta(t)$, *l'équation*

$$\dot{\beta} = (g,\ f) \tag{4.38}$$

*et la deuxième des conditions* (4.35), *alors quel que soit* $t$ *la fonction vectorielle* $y(t)$ *vérifie la condition*

$$(g(t),\ y(t)) = \beta(t). \tag{4.39}$$

D'après ce théorème, toute condition linéaire (4.34) peut être transférée d'un point $t_i$ en un point arbitraire. Il suffit pour cela de

résoudre un problème de Cauchy pour le système adjoint et un autre pour l'équation scalaire (4.38).

REMARQUE. Nous avons examiné une procédure de transfert des conditions aux limites pour des équations différentielles linéaires. En analyse numérique, cette procédure est appelée aussi *méthode de factorisation*. Nous utiliserons le terme de factorisation dans la suite de l'exposé.

Ce résultat ouvre la voie à la standardisation des méthodes de résolution des problèmes d'optimisation linéaires à fonctionnelle quadratique. La première étape de la résolution consiste en une réduction du problème linéaire à un problème aux limites pour le système linéaire (4.32). C'est une procédure courante qui met en jeu une inversion et une multiplication de matrices. La deuxième étape est le transfert des conditions aux limites qui sont données par les conditions de transversalité aux extrémités de la trajectoire. Cette étape est classique aussi : elle implique la résolution d'un problème de Cauchy. Après le transfert des conditions aux limites, il reste à résoudre un ordinaire problème de Cauchy. Donc, le problème linéaire de commande optimale à fonctionnelle quadratique se ramène à la résolution de trois problèmes de Cauchy, c'est-à-dire à une suite de procédures classiques.

Certes ce problème a ses propres écueils qui en principe sont liés aussi à l'instabilité. Plus, les remarques faites en discutant la méthode d'ajustage nous permettent d'affirmer que nous aurons toujours affaire à l'instabilité en procédant au transfert des conditions aux limites. Donc, la réalisation de la méthode de factorisation sur ordinateur dans la forme dans laquelle elle a été développée est en général aussi compliquée que celle de la méthode d'ajustage.

Cependant la méthode de factorisation possède la remarquable propriété d'être toujours modifiée de sorte que la procédure de calcul soit stable. En effet, le transfert des conditions aux limites peut être effectué d'une infinité de manières parmi lesquelles on peut toujours trouver des stables. Voyons comment il faut s'y prendre. Le transfert des conditions aux limites sera réalisé par un prolongement analytique de $l_i$ et $\alpha_i$, autrement dit, on construira des fonctions différentiables $g$ et $\beta$ qui se transforment respectivement en $l_i$ et $\alpha_i$ pour $t = t_i$ et satisfont la condition (4.39) à tout instant $t$. Ce prolongement n'étant pas unique, au lieu de la fonction trouvée $g(t)$ on peut en prendre une autre, $\hat{g}(t)$, telle que son module soit une constante donnée. Posons

$$\hat{g}(t) = m(t)\, g(t),$$

où $g(t)$ est une fonction vectorielle satisfaisant l'équation adjointe (4.36) et les conditions (4.35), $m(t)$, une fonction à déterminer.

Formons l'équation que doit satisfaire la fonction $\hat{g}(t)$:

$$\dot{\hat{g}} = \dot{m}g + m\dot{g} = \dot{m}g - mA^*g,$$

or $g = \hat{g}/m$, donc

$$\dot{\hat{g}} = \frac{\dot{m}}{m}\hat{g} - A^*\hat{g}. \tag{4.40}$$

Choisissons maintenant la fonction $m(t)$ de telle sorte que le module de la fonction $\hat{g}$ soit une constante, par exemple: $(\hat{g}, \hat{g}) = 1$. A cet effet il est nécessaire et suffisant (dans le cas où $(l_i, l_i) = 1$) que $(\dot{\hat{g}}, \hat{g}) = 0$. Cette condition nous donne une équation pour $m(t)$:

$$\frac{\dot{m}}{m}(\hat{g}, \hat{g}) - (\hat{g}, A\hat{g}) = 0,$$

d'où

$$\frac{\dot{m}}{m} = \frac{(A^*\hat{g}, \hat{g})}{(\hat{g}, \hat{g})}.$$

En portant cette expression dans l'équation (4.40), on trouve une équation pour $\hat{g}$:

$$\dot{\hat{g}} = -A^*\hat{g} + \frac{(A^*\hat{g}, \hat{g})}{(\hat{g}, \hat{g})}\hat{g}. \tag{4.41}$$

Soumettons la fonction $\hat{g}(t)$ à la condition subsidiaire $\hat{g}(t_i) = l_i$. Il nous reste seulement à déterminer la nouvelle fonction $\beta(t)$. Posons

$$\beta(t) = (\hat{g}(t), y(t))$$

et cherchons

$$\dot{\beta} = (\hat{g}, f) + \frac{(A^*\hat{g}, \hat{g})}{(\hat{g}, \hat{g})}\beta. \tag{4.42}$$

Donc, le transfert de la condition aux limites s'opère de nouveau à l'aide de la formule $(\hat{g}, y) = \beta$, où $\hat{g}(t)$ est définie comme la solution du problème de Cauchy $\hat{g}(t_i) = l_i$ pour l'équation (4.41) et la fonction $\beta(t)$, à partir de l'équation (4.42) et de la condition $\beta(t_i) = \alpha_i$.

Ce résultat est très important: le transfert des conditions aux limites étant toujours stable, nous devons le réaliser de telle manière que le problème de Cauchy pour l'équation (4.33) le soit aussi. Si par exemple nous avons de nouveau un problème à extrémités fixes

($x(0) = x_0$, $x(T) = x_T$) et si la solution de l'équation $\dot{x} = Ax$ est stable, alors les conditions aux limites doivent être transférées de l'extrémité droite de la trajectoire vers l'extrémité gauche. Si la solution est instable, alors il faut transférer les conditions aux limites de l'extrémité gauche vers la droite et résoudre le problème de Cauchy de la droite vers la gauche.

On constate donc que l'appareil développé est un outil très souple de résolution des problèmes aux limites et nous avons presque toujours la possibilité de tourner les difficultés apportées par l'instabilité du calcul.

REMARQUES. 1. Les conditions aux limites transférées deviennent

$$(\hat{g}(t_i),\ y(t_i)) = \beta(t_i).$$

Pour pouvoir utiliser les méthodes standards, il nous faut encore résoudre ce système par rapport à la fonction vectorielle $y(t_i)$. Or la matrice de ce système peut être mal conditionnée. Si tel est le cas, on surmonte cette difficulté par une généralisation appropriée de la méthode exposée ci-dessus (cf. [6]).

2. L'application des équations adjointes à l'analyse des systèmes d'équations différentielles linéaires était connue déjà de Legendre et de Liouville. Donc, la possibilité de transférer les conditions aux limites d'un point à un autre sans résoudre de problème aux limites remonte au premier tiers du XIXe siècle (la procédure s'appelait méthode de factorisation). Mais le problème de la stabilité de cette procédure ne s'est posé naturellement qu'après l'apparition des calculatrices. La solution complète de ce problème a été donnée par A. Abramov (cf. [16]). Notre exposé de la procédure de transfert s'inspire des idées d'Abramov.

Donc, la résolution des problèmes linéaires de commande optimale à fonctionnelle quadratique (sans contraintes sur la commande) peut être ramenée à un système de procédures régulières. Cette circonstance peut être utilisée à la construction de méthodes itératives de divers type.

Supposons maintenant que l'équation régissant le système commandé n'est pas linéaire et est de la forme

$$\dot{x} = f(x, u). \tag{4.43}$$

Posons de nouveau le problème de commande optimale:

$$x(0) = x_0, \quad x(T) = x_T, \tag{4.44}$$

$$J(u) = \int_0^T F(x, u)\, dt \Rightarrow \min \tag{4.45}$$

et supposons que la commande n'est soumise à aucune contrainte et que l'horizon $T$ est fixe.

Donnons-nous une commande $u = u^0(t)$ qui sera traitée comme une approximation initiale de la solution. En portant cette commande dans l'équation (4.43) et en résolvant le problème de Cauchy $x(0) = x_0$ pour cette équation, on trouve une trajectoire $x = x^0(t)$. Cette trajectoire ne satisfera en général pas les conditions à l'extrémité droite, c'est-à-dire $x^0(T) \neq x_T$. A la commande $u^0(t)$ sera associée une certaine valeur de la fonctionnelle $J_0$.

Introduisons maintenant les nouvelles variables $x = x^0 + y$, $u = u^0 + v$. En supposant $y$ et $v$ petits et en ne retenant dans l'équation (4.43) que les termes linéaires en $y$ et $v$, on obtient le système d'équations linéaires

$$\dot{y} = Ay + Bv, \tag{4.46}$$

où $A$ et $B$ sont les matrices

$$A = \left(\frac{\partial f}{\partial x}\right)_{\substack{x=x^0 \\ u=u^0}}, \quad B = \left(\frac{\partial f}{\partial u}\right)_{\substack{x=x^0 \\ u=u^0}}.$$

La fonction $y(t)$ satisfait des conditions aux limites nulles à l'extrémité gauche. A l'extrémité droite imposons-lui les conditions

$$y(T) = y_T = x_T - x^0(T). \tag{4.47}$$

Faisons un changement de variables dans la fonctionnelle et gardons non seulement les termes linéaires en $y$ et $v$ mais aussi les termes quadratiques. Nous obtenons ainsi la nouvelle fonctionnelle

$$\varphi = \int_0^T [(\alpha, y) + (\beta, v) + (Cy, y) + (Dv, y) + (Ev, v)]\, dt. \tag{4.48}$$

La signification des notations est évidente: $\alpha$ et $\beta$ sont les vecteurs $\alpha = \partial F/\partial x$, $\beta = \partial F/\partial u$, $C$, $D$ et $E$ sont les matrices

$$C = \frac{1}{2}\left(\frac{\partial^2 F}{\partial x^i\, \partial x^j}\right), \quad i, j = 1, \ldots, n;$$

$$D = \frac{1}{2}\left(\frac{\partial^2 F}{\partial x^i\, \partial u^j}\right), \quad i = 1, \ldots, n; \quad j = 1, \ldots, m;$$

$$E = \frac{1}{2}\left(\frac{\partial^2 F}{\partial u^i\, \partial u^j}\right), \quad i, j = 1, \ldots, m.$$

Toutes ces quantités sont calculées pour $x = x^0(t)$, $u = u^0(t)$.

Nous sommes ainsi conduits à la recherche d'une commande $v$ transférant le système (4.46) de l'origine des coordonnées à l'état (4.47) pendant un intervalle de temps $T$ et réalisant le minimum de la fonctionnelle (4.48). Nous avons vu que ce problème se ramenait à un problème aux limites pour un système linéaire, c'est-à-dire qu'il peut être résolu par des méthodes régulières. Nous obtenons en

définitive une nouvelle équation

$$u_1 = u^0 + v \tag{4.49}$$

et une nouvelle trajectoire de phase $x_1$ qui sera solution du problème de Cauchy pour le système initial (4.43) avec une commande égale à $u_1$.

Calculons la nouvelle valeur de la fonctionnelle $J_1 = J_1(x_1, u_1)$. Si $J_1 < J_0$, c'est que la solution $(u_1, x_1)$ améliore l'approximation initiale et l'on peut répéter cette procédure en prenant $x_1$ et $u_1$ pour approximation initiale.

Si $J_1 \geqslant J_0$, il faut procéder comme dans la réalisation de la méthode de Newton ou de la méthode de plus grande pente, c'est-à-dire prendre

$$u_1 = u^0 + kv,$$

où $k \in ]0, 1[$. Ceci étant, la fonctionnelle $J_1$ devient une fonction de $k$. On cherche ensuite un $k$ pour lequel $J_1(k) \underset{k}{\Rightarrow} \min$.

La convergence d'un tel schéma itératif n'a pas été étudiée, mais les innombrables problèmes résolus attestent de son efficacité, pourvu que l'approximation initiale soit « assez bonne ».

REMARQUE. A la fin des années soixante R. Bellman proposa une méthode itérative de résolution des problèmes de commande optimale qui fut appelée méthode de quasi-linéarisation (cf. [1]). Cette méthode est pratiquement identique à la méthode de linéarisation exposée plus haut. Cependant elle en diffère par le fait que Bellman n'utilise pas la technique de transfert des conditions aux limites qui assure la stabilité du calcul.

La méthode de factorisation et les schémas itératifs dans lesquels elle intervient sont performants dans la résolution de problèmes sans contraintes sur la commande. Cette méthode est également utilisée dans des cas plus généraux et notamment à la résolution de problèmes aux limites non linéaires. Soit donc le système d'équations non linéaire

$$\dot{x} = \varphi(x), \tag{4.50}$$

où $x$ est un vecteur de dimension paire, $\varphi(x)$, une fonction vectorielle non linéaire. Supposons que la moitié des conditions est donnée à l'extrémité gauche et l'autre moitié à l'extrémité droite de la trajectoire.

Le schéma itératif utilisant la méthode de factorisation consiste en ce qui suit. Définissons par un procédé quelconque l'approximation initiale $x^0$ et mettons l'équation (4.50) sous la forme

$$\dot{x} = A(x^0)\,x + L(x, x^0),$$

où $L(x, x^0) = \varphi(x) - A(x^0)\,x$. Si $\varphi(x)$ est une fonction différen-

tiable, alors l'opérateur $A$ est la matrice des dérivées partielles

$$A = \left(\frac{\partial \varphi^i}{\partial x^j}\right)_{x=x^0}, \quad i, j = 1, \ldots, n,$$

et la structure du schéma itératif est évidente:

$$\dot{x}_k = A(x_{k-1})\, x_k + L(x_{k-1}, x_{k-1}).$$

Ces schémas de calcul revêtent un caractère euristique pur; il existe cependant suffisamment de processus pour lesquels ils ont permis d'obtenir des résultats utiles.

REMARQUE. Dans ce numéro, on a exposé la méthode de transfert des conditions aux limites pour des équations linéaires qui conduit toujours à des procédures de calcul stables. Cette méthode permet notamment d'acquérir la solution numérique du problème de Cauchy pour une équation de la forme

$$\dot{x} = A(t)\, x + f(t)$$

même lorsque les solutions particulières de cette équation sont des fonctions à croissance rapide. Supposons que les conditions initiales sont de la forme

$$x(0) = x_0 \qquad (x^i(0) = x_0^i, \quad i = 1, \ldots, k),$$

où $x_0^i$ est la $i$-ème coordonnée du vecteur $x_0$. Ces conditions peuvent être récrites sous la forme

$$(e_i, x) = x_0^i,$$

où $e_i$ est le vecteur unitaire du $i$-ème axe.

Supposons maintenant que l'on ait à calculer $x(T)$. Il nous faut pour cela résoudre le problème de Cauchy formulé plus haut et intégrer numériquement l'équation entre $t = 0$ et $t = T$. Mais si parmi les valeurs propres de la matrice $A(0)$ il en est qui possèdent une partie réelle positive élevée, alors les solutions particulières de l'équation $\dot{x} = A(t)\, x$ seront à croissance rapide et toute méthode de résolution numérique sera très difficile à réaliser. Mais pour calculer $x(T)$ on peut éviter le problème de Cauchy s'il est instable. A sa place il faut résoudre $n$ problèmes de Cauchy stables

$$\hat{g}_i(0) = e_i,$$

qui nous donneront $n$ relations

$$(\hat{g}_i(T), x(T)) = \beta_i(T).$$

Les approches exposées permettent d'améliorer de nombreuses procédures numériques et notamment le caractère des procédures itératives dans la méthode d'ajustage.

c) *Méthodes utilisant une procédure de résolution de problèmes avec une extrémité libre.* Les problèmes à extrémité droite libre jouissent d'une remarquable propriété: pour obtenir la solution exacte d'un problème de commande optimale si celle-ci est linéaire en les variables de phase, il suffit de résoudre deux problèmes de Cauchy.

Considérons le système

$$\dot{x} = A(t)\, x + \varphi(t, u), \tag{4.51}$$

où $\varphi$ est une fonction non linéaire arbitraire et la commande $u(t)$ satisfait des contraintes de la forme

$$u \in G. \tag{4.52}$$

L'état initial du système est donné:

$$x(0) = x_0, \tag{4.53}$$

quant à l'extrémité droite de la trajectoire, elle n'est soumise à aucune contrainte.

Posons pour ce système le problème

$$J = \int_0^T F(x, u)\, dt \Rightarrow \min, \tag{4.54}$$

où $F$ est une fonction linéaire en $x$ de la forme $F(x, u) = (c, x) + f(u)$.

D'après ce qui a été dit au début du paragraphe, pour ramener ce problème à un problème aux limites il faut composer le hamiltonien

$$H = (\psi, Ax) + (\psi, \varphi(t, u)) - (c, x) - f(u) \tag{4.55}$$

et l'équation pour les variables adjointes (les impulsions)

$$\dot{\psi} = -\frac{\partial H}{\partial x} = -A^*\psi + c. \tag{4.56}$$

L'extrémité droite de la trajectoire n'étant soumise à aucune contrainte, la condition de transversalité (4.16) nous donne

$$\psi(T) = 0.$$

On remarquera que l'équation pour les impulsions ne contient pas dans ce cas la variable de phase et peut être intégrée indépendamment de l'équation (4.51). Les valeurs des impulsions à l'extrémité droite de la trajectoire étant connues, la résolution du problème de Cauchy (4.16), (4.56) de droite à gauche nous donne les impulsions $\psi(t)$ indépendamment de $x$.

On peut de même déterminer la commande $u(t)$ indépendamment de $x$. En effet, la commande se déduit à partir de la condition $H \Rightarrow \max$, mais le hamiltonien $H$, ainsi qu'il résulte de (4.55), ne renferme que deux termes dans lesquels figure la commande: le produit scalaire $(\psi, \varphi(t, u))$ et $f(u)$. Donc, la commande $u(t)$ se détermine à partir de la condition

$$-f(u) + (\psi, \varphi(t, u)) \Rightarrow \max_{u \in G}. \tag{4.57}$$

De là on tire immédiatement $u = u(t)$, puisqu'on a déjà trouvé $\psi(t)$ comme fonction de $t$. Une fois qu'on a la commande $u(t)$, on

la porte dans l'équation (4.51) et en résolvant le problème de Cauchy (4.53), on détermine la trajectoire de phase $x(t)$.

Ainsi, le problème de commande optimale à une extrémité libre, linéaire en la variable de phase, peut être résolu par des méthodes régulières, c'est-à-dire ramené à deux problèmes de Cauchy et un problème de programmation non linéaire (4.57).

Cette circonstance a suggéré la mise au point de diverses procédures itératives. Le premier travail de cette nature a été publié par L. Chatrovski [23] en U.R.S.S. et par A. Brasson aux U.S.A. Ces deux auteurs ont proposé des procédures itératives pratiquement identiques appelées en U.R.S.S. méthode de Chatrovski-Brasson. Ces procédures s'appuient sur les raisonnements suivants.

Supposons que nous ayons un problème général de commande optimale (4.43), (4.45), (4.52), (4.53) avec l'extrémité droite libre. Donnons-nous une commande initiale $u^0$ et calculons la trajectoire $x^0$ associée. Faisons ensuite le changement de variables

$$x = x^0 + y, \qquad u = u^0 + v$$

dans l'équation (4.43) et dans la fonctionnelle (4.45) et gardons les termes linéaires en $y$ et $v$. On obtient ainsi l'équation pour la variable de phase

$$\dot{y} = Ay + Bv \tag{4.58}$$

et la fonctionnelle

$$J_1 = \int_0^T [(c, y) + (f, v)]\, dt, \tag{4.59}$$

où

$$c = \left(\frac{\partial F}{\partial x}\right)_{y=0}; \quad f = \left(\frac{\partial F}{\partial u}\right)_{v=0}.$$

En résolvant le problème $J_1 \Rightarrow \min$ sous la condition (4.58) et la contrainte $u^0 + v \in G$, on obtient une nouvelle commande. On prend ensuite cette commande pour approximation et on recommence les calculs.

Cette procédure de linéarisation n'est pas assez payante sur le plan de sa réalisation sur ordinateur. En effet, dans la résolution numérique du problème de Cauchy, le passage à un système linéaire ne se traduit par aucun gain. Pour la programmation il est important d'avoir une forme condensée d'écriture. La linéarisation implique le calcul des dérivées et l'on perd sur le plan de la condensation de l'écriture de l'équation non linéaire. I. Krylov et F. Tchernoousko [48] ont proposé leur propre modification de cette méthode, modification qui se ramène aux procédures suivantes.

7—01523

1) On se donne une approximation initiale $u^0(t)$, puis l'on résout le problème de Cauchy pour l'équation (4.43) et l'on détermine ensuite $x^0(T)$.

2) Les valeurs des impulsions étant données à l'extrémité droite de la trajectoire (cf. (4.16)), on peut résoudre le problème de Cauchy

$$\psi(T) = 0$$

pour le système

$$\dot{\psi}_i = -\sum_{j=1}^{n} \frac{\partial f^j}{\partial x^i}(x^0, u^0, t)\,\psi_j + \frac{\partial F}{\partial x^i}(x^0, u^0), \quad i = 1, \ldots, n. \quad (4.60)$$

Le problème est résolu de droite à gauche, c'est-à-dire de $t = T$ vers $t = 0$.

3) On intègre ensuite le système (4.60) pour trouver une commande $u^1$ telle que

$$H = (\psi, f)_{x=x^0} - F(x^0, u) \Rightarrow \max_{u \in G}.$$

4) On répète ensuite cette procédure en prenant $u^1$ pour approximation initiale.

La méthode de Krylov-Tchernoousko est bien plus économique que celle de Chatrovski-Brasson et plus commode à réaliser sur machine. Mais dans le cas général elle est divergente. Pour en améliorer la convergence, on peut par exemple remplacer la procédure d'intégration du système (4.43) avec la commande $u_1$ par l'intégration avec la commande

$$\tilde{u}_1 = u^0 + \frac{u_1 - u^0}{ku^0},$$

où $k$ est déduit à partir de la condition $J(\tilde{u}_1) < J(u^0)$. Cette méthode a suscité de nombreuses recherches et les conditions de sa convergence ont été établies par M. Beïko et I. Beïko [20].

Les méthodes de calcul des programmes optimaux utilisant les conditions nécessaires sous forme du principe du maximum de Pontriaguine constituent un chapitre autonome de l'analyse numérique. Dans ce paragraphe, nous nous sommes bornés à un exposé schématique de quelques idées seulement illustrant les principales orientations de ces méthodes. Au chapitre VI nous reviendrons sur les problèmes de détermination effective des commandes optimales et développerons quelques principes simplifiant les procédures numériques.

## § 5. Problème de rapidité

Les problèmes de rapidité occupent une place particulière en théorie des systèmes commandés. La position formelle du problème est la suivante. Soit un système commandé

$$\dot{x} = f(x, u, t), \tag{5.1}$$

dont l'état initial est fixé:

$$x(0) = x_0. \tag{5.2}$$

On demande de trouver une commande $u(t)$ vérifiant la condition

$$u(t) \in G_u \tag{5.3}$$

et faisant passer le système dans un ensemble terminal $G_T$ pendant un intervalle de temps $T$ minimal. L'ensemble $G_T$ est très général. Dans ce problème, la fonctionnelle est la durée $T$ du processus de commande.

Les problèmes de rapidité ne se rencontrent pas fréquemment dans leur position classique. Dans les projets, il y a d'autres facteurs qui priment sur le temps: les dépenses en argent, en matériel, etc. Mais les problèmes de rapidité ont leur raison d'être: le temps de réalisation d'un projet est aussi un critère important. Il illustre souvent la « réalisabilité » d'un projet, son adéquation. Le problème de rapidité permet de fixer les délais limites de réalisation d'un projet (en fonction des investissements), de placement d'un système économique sur une courbe d'évolution donnée, etc. Les méthodes de recherche de commandes qui soient optimales du point de vue de la rapidité sont d'une grande importance pratique et dans le chapitre suivant nous les situerons dans le système général des procédures de la méthode de programmation.

Les problèmes de rapidité ont fortement contribué au développement de la théorie de la commande. C'est pour eux que le principe du maximum a été initialement formulé. On verra plus bas que formellement les problèmes de rapidité sont un cas particulier des problèmes de commande et les conditions nécessaires que doit satisfaire la commande dans les problèmes de rapidité résultent facilement de généralités de la théorie des systèmes commandés et du principe du maximum. Les problèmes de rapidité méritent cependant d'être spécialement étudiés pour leurs particularités sur le plan du calcul numérique.

Pour simplifier les raisonnements, on admettra que l'ensemble terminal est constitué d'un seul point, c'est-à-dire que la position finale du système est fixée:

$$x(T) = x_T.$$

Considérons le problème général de commande optimale avec une fonctionnelle intégrale: déterminer une trajectoire $x(t)$ du système (5.1) transférant le système de l'état (5.2) à l'état (5.4) et telle que le long d'elle la fonctionnelle

$$J(u)=\int_0^T F(x,\ u,\ t)\,dt \tag{5.5}$$

prenne sa valeur minimale sous la condition (5.3).

Ecrivons les conditions nécessaires sous forme du principe du maximum. Formons à cet effet le hamiltonien

$$H=(\psi,\ f)-F(x,\ u,\ t), \tag{5.6}$$

où les impulsions $\psi$ satisfont le système d'équations

$$\dot{\psi}=-\frac{\partial H}{\partial x}=-f_x^*\psi+F_x. \tag{5.7}$$

Le principe du maximum nous dit qu'une condition nécessaire pour qu'une fonction $u(t)$ soit commande optimale dans le problème (5.1) à (5.5) est qu'à tout instant elle maximise le hamiltonien

$$H(x,\ \psi,\ u,\ t)\Rightarrow \max_{u\in G_u}. \tag{5.8}$$

Le temps $T$ peut être fixé ou libre, c'est-à-dire déterminé par la résolution du problème. Mais si le temps n'est pas donné, il faut poser encore une condition pour le trouver. Signalons que si le temps est fixé, on obtient un problème aux limites qui nous permet de déterminer la commande et la trajectoire de phase. En modifiant $T$, on aura des commandes et des trajectoires différentes.

Il existe plusieurs façons de se donner la condition subsidiaire pour la détermination de $T$. Cette condition peut par exemple résulter des conditions de transversalité énoncées dans le paragraphe précédent: si aucune contrainte n'est imposée à la variable de phase $x^i(t)$ à l'instant initial $t=0$ ou final $t=T$, alors la variable adjointe correspondante — l'impulsion $\psi_i(t)$ — doit s'annuler pour $t=0$ ou pour $t=T$, c'est-à-dire

$$\psi_i(0)=0 \tag{5.9}$$

ou

$$\psi_i(T)=0. \tag{5.9'}$$

Introduisons maintenant une variable de phase supplémentaire $x^{n+1}$ à l'aide de l'égalité $\dot{x}^{n+1}=0$ et une variable $\tau\in[0,\ 1]$. Faisons le changement de variables $t=x^{n+1}\tau$. On a alors $x^{n+1}=T$.

Ceci nous permet de mettre le système (5.1) à (5.5) sous la forme

$$\frac{dx}{d\tau}=x^{n+1}f(x,\ u,\ x^{n+1}\tau),\qquad \frac{dx^{n+1}}{d\tau}=0, \tag{5.10}$$

$$J(u)=x^{n+1}\int_0^1 F(x,\ u,\ x^{n+1}\tau)\,d\tau. \tag{5.11}$$

Composons les conditions du principe du maximum pour le problème (5.10), (5.11). Introduisons le hamiltonien $H^*$ pour le système (5.10):

$$H^* = x^{n+1}\cdot H + \psi_{n+1}\cdot 0.$$

Les impulsions $\psi_1, \ldots, \psi_n, \psi_{n+1}$ satisferont les conditions

$$\frac{d\psi_i}{d\tau}=-\frac{\partial H^*}{\partial x^i}=-x^{n+1}\frac{\partial H}{\partial x^i},\quad i=1,\ \ldots,\ n, \tag{5.12}$$

$$\frac{d\psi_{n+1}}{d\tau}=-\frac{\partial H^*}{\partial x^{n+1}}=-H-x^{n+1}\frac{\partial H}{\partial t}\frac{dt}{dx^{n+1}}=$$

$$=-H-\frac{\partial H}{\partial t}\tau x_{n+1}. \tag{5.13}$$

La variable de phase $x^{n+1}$ n'étant soumise à aucune contrainte pour $t = 0$ et $t = T$, les conditions de transversalité s'écrivent

$$\psi_{n+1}(0) = \psi_{n+1}(1) = 0,$$

ou encore

$$\int_0^1 \left[-H-\frac{\partial H}{\partial t}\tau x^{n+1}\right]d\tau=0. \tag{5.14}$$

L'expression (5.14) est la condition subsidiaire cherchée qui permet de déterminer la constante inconnue $x^{n+1}$.

Le problème de rapidité est un cas particulier d'un problème à horizon non fixe avec $F(x, u, t) = 1$. Ceci étant, $H = (\psi, f) - 1$ et le problème de rapidité peut s'énoncer sous la forme suivante: trouver la solution du système

$$\frac{dx}{d\tau}=x^{n+1}f(x,\ u,\ x^{n+1}\tau),\qquad \frac{d\psi}{d\tau}=-x^{n+1}f_x^*\psi, \tag{5.15}$$

avec les conditions aux limites

$$x(0) = x_0,\qquad x(1) = x_T, \tag{5.16}$$

où la commande $u$ satisfait le principe du maximum

$$H=(\psi,\ f)\,x_{n+1}\Rightarrow \max_u,$$

et le paramètre $x^{n+1}$ est déduit de la condition

$$\int_0^1 \left[(\psi, f) + \frac{\partial H}{\partial t} \tau x^{n+1}\right] d\tau = 1.$$

Le problème (5.15), (5.16) possède de nombreux traits spécifiques du point de vue de l'organisation des calculs: en particulier, la présence du paramètre $x^{n+1}$ qui doit être déterminé à partir de la condition subsidiaire (5.14) complique la résolution numérique. On dispose de plusieurs méthodes pour tourner ces difficultés. Indiquons quelques-unes d'entre elles.

a) *Utilisation des méthodes de la programmation non linéaire.* Si l'horizon $T$ n'est pas trop élevé (c'est-à-dire si la précision du problème n'implique pas une subdivision trop fine de l'intervalle temporel), alors le passage à une approximation aux différences du problème et son remplacement par un problème de programmation non linéaire nous permettent d'utiliser la bibliothèque des programmes standards.

Le problème de programmation non linéaire qui se présente ici se formule comme suit: minimiser la fonction linéaire

$$x^{n+1} \Rightarrow \min$$

sous les conditions

$$x(i+1) = x(i) + \Delta x^{n+1} f(x(i), u_i, x^{n+1}\tau_i),$$

$$x(0) = x_0,\ x(N) = x_T,\ u_i \in U_i,\ \Delta = 1/N,\ \tau_i = i\Delta.$$

b) *Passage à une coordonnée monotone.* Dans de nombreux problèmes techniques, on peut dégager une coordonnée monotone. Dans les problèmes portant sur le mouvement des satellites artificiels, cette coordonnée est toujours l'angle polaire $\varphi$. Le changement de la variable indépendante $t$ en la variable monotone $\varphi$ ramène alors le problème à horizon variable à un problème classique de commande optimale.

c) *Introduction d'une fonctionnelle subsidiaire.* Fixons la variable $x^{n+1}$ en la prenant $< T_{\min}$ et résolvons le problème à horizon fixe de minimisation de la fonctionnelle

$$J^* = ((x(1) - x_T),\quad R(x(1) - x_T)), \tag{5.17}$$

où $R$ est une matrice définie positive caractérisant la précision de réalisation de l'objectif de la commande.

En résolvant successivement les problèmes de minimisation de la fonctionnelle (5.17) pour $x_0^{n+1} < x_1^{n+1} < x_2^{n+1}$, on trouve les quantités

$$J^*(x^{n+1}).$$

Ces quantités expriment la dépendance de la fonctionnelle (5.17) par rapport à la variable $x^{n+1}$. En approchant cette dépendance par une fonction assez simple, on trouve la solution de l'équation

$$J^* (x^{n+1}) = 0, \tag{5.18}$$

après quoi on résout le problème aux limites (5.15), (5.16) pour la valeur $x^{n+1}$ déduite de (5.18).

## § 6. Méthodes directes de calcul des programmes optimaux

a) *Deux méthodes de réduction des problèmes de commande optimale à des problèmes d'optimisation finidimensionnelle.* Le passage à une description finidimensionnelle (discrète) des problèmes continus ouvre la voie à l'utilisation de l'appareil élaboré de la programmation non linéaire et dynamique. Dans le paragraphe précédent, nous nous sommes déjà servis de l'approximation finidimensionnelle. Revenons à cette question. Considérons le système

$$\dot{x} = f(x, u) \tag{6.1}$$

et la fonctionnelle

$$J(u) = \int_0^T F(x, u)\, dt. \tag{6.2}$$

Limitons-nous à un schéma aux différences élémentaire et remplaçons (6.1) et (6.2) par les expressions suivantes:

$$x_{i+1} = x_i + \tau f(x_i, u_i), \tag{6.3}$$

$$J = \tau \sum_{i=0}^{N-1} F(x_i, u_i) \quad i = 0, 1, \ldots, N-1, \tag{6.4}$$

où

$$t_i = i\tau, \quad x(t_i) = x_i, \; u(t_i) = u_i.$$

Ajoutons aux équations (6.3) la condition initiale

$$x(0) = x_0 \tag{6.5}$$

et la condition à l'extrémité droite que, pour fixer les idées, nous prendrons de la forme

$$x(T) = x_T = x_N, \tag{6.6}$$

où $x_N$ est un vecteur fixé. Le vecteur de phase et la commande peuvent être de plus soumis à d'autres contraintes.

Ce changement ramène le problème de minimisation de la fonctionnelle (6.2) à la minimisation de la fonction (6.4) sous les con-

ditions (6.3), (6.5) et (6.6). Nous sommes ainsi arrivés à un problème de programmation non linéaire. Mais le caractère des liaisons (6.3) confère à ce problème des traits particuliers qui impliquent la mise en œuvre de méthodes spéciales de résolution.

L'égalité (6.3) permet d'éliminer successivement les vecteurs de phase:

$$x_1 = x_0 + \tau f(x_0, u_0) = \Phi_1(u_0),$$

$$x_2 = \Phi_1(u_0) + \tau f(\Phi_1(u_0), u_1) = \Phi_2(u_0, u_1),$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$x_k = \Phi_{k-1}(u_0, \ldots, u_{k-2}) + \tau f(\Phi_{k-1}(u_0, \ldots, u_{k-2}), u_{k-1}) = \\ = \Phi_k(u_0, \ldots, u_{k-1}),$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

La fonctionnelle (6.4) devient une fonction des seuls vecteurs $u_0, \ldots, u_{N-1}$:

$$J = \sum_{i=0}^{N-1} I_i(u_0, u_1, \ldots, u_i), \tag{6.7}$$

où

$$I_i(u_0, u_1, \ldots, u_i) = \tau F(\Phi_i(u_0, \ldots, u_{i-1}), u_i).$$

On voit donc que la fonctionnelle $J$ s'est transformée en une somme de termes $I_i$ tels que les termes d'indice $i$ ne dépendent que des $i+1$ premières variables inconnues. Les fonctions (6.7) seront appelées fonctions à introduction progressive des variables.

Un autre procédé de réduction d'un problème de commande optimale à un problème d'optimisation finidimensionnelle est lié à l'opération dite élémentaire. Supposons que dans l'espace des états $(x, t)$ l'on ait mené des plans $t = t_i = i\tau$, désignés par $\Sigma_i$. Ces plans sont coupés aux points $x_i$ par la trajectoire $\gamma$ du système. Introduisons maintenant un opérateur $B(x_i, x_{i+1})$ qui à tout couple de points $x_i$ et $x_{i+1}$ associe la commande $u_i$ qui pendant un intervalle de temps $\tau$ fait passer le système de l'état $x_i$ à l'état $x_{i+1}$, et la portion de trajectoire $\gamma_{i,i+1}$ joignant ces points. Nous noterons ceci sous la forme

$$(\gamma_{i,i+1}, u_i) = B(x_i, x_{i+1})$$

et nous appellerons l'opérateur $B(x_i, x_{i+1})$ *opération élémentaire*. La fonctionnelle $J$ peut maintenant être mise sous la forme

$$J(x, u) = \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} F(\gamma_{i,i+1}, u_i)\, dt = \sum_{i=0}^{N-1} \varphi_i(x_i, x_{i+1}) \tag{6.8}$$

Si donc nous est donnée une opération élémentaire $B(x_i, x_{i+1})$, alors la trajectoire est définie par un nombre fini de points $x_i$, intersections de cette trajectoire avec les plans $\Sigma_i$.

La notion d'opération élémentaire peut être généralisée. Nous ne rattacherons pas la construction de l'arc $\gamma_{i,i+1}$ à un segment de trajectoire de phase. Définissons l'opération $B(x_i, x_{i+1})$ comme une procédure de construction d'un vecteur $u_i$ et d'un segment $\gamma_{i,i+1}$ joignant des points donnés $x_i$ et $x_{i+1}$.

Cette opération nous permet d'approcher la trajectoire de phase par une ligne polygonale composée d'arcs $\gamma_{i,i+1}$ et de ramener le problème initial de commande optimale à la détermination du minimum d'une fonction d'un nombre fini de variables (6.8). La ligne polygonale sera appelée *ligne polygonale d'Euler.*

Si pour « longueur » du segment $\gamma_{i,i+1}$ on prend

$$\Delta J = \int_{t_i}^{t_{i+1}} F(\gamma_{i,\,i+1},\ u_i)\,dt.$$

on peut encore formuler le problème initial dans les termes suivants: parmi les lignes polygonales d'Euler joignant deux points donnés $x_0$ et $x_T$, trouver la plus courte.

La construction de l'opération élémentaire est toujours assez compliquée et l'on ne dispose d'aucune méthode classique pour le faire. Parfois on l'obtient très facilement, par exemple dans le problème variationnel simple où les équations (6.1) sont de la forme $\dot{x} = u$. On peut alors poser $u_i = (x_{i+1} - x_i)/\tau$.

Si les vecteurs $u$ et $x$ sont de même dimension, la construction de l'opération élémentaire peut être ramenée à la résolution de l'équation transcendante

$$\frac{x_{i+1} - x_i}{\tau} = f(x_i,\ u_i).$$

Si le vecteur $u$ est de dimension inférieure à celle du vecteur de phase $x$ (cas type), la construction de l'opération élémentaire se complique singulièrement. Dans ce cas, il est plus commode de chercher une commande constante $u_i$ en la choisissant telle que le point $x^*_{i+1}$ défini par

$$x^*_{i+1} = x_i + \tau f(x_i,\ u_i),$$

soit le plus proche de $x_{i+1}$ (fig. 2.1).

Fig. 2.1

REMARQUE. La procédure de choix de la commande $u_i$ doit être conduite de telle sorte que l'écrat $\delta_i(\tau) = \| x^*_{i+1} - x_{i+1} \| \to 0$ avec $\tau$.

b) *Méthode du gradient.* Les méthodes classiques de résolution numérique des problèmes de programmation non linéaire obtenus par discrétisation des problèmes de commande optimale se simplifient légèrement en raison de la forme spécifique des contraintes. Illustrons ceci sur l'exemple de la méthode du gradient.

Soit à minimiser la fonction

$$J = \sum_{i=0}^{N-1} I_i (u_0, u_1, \ldots, u_i), \tag{6.9}$$

où $u_i$ sont des vecteurs de dimension $m$. Chaque pas de la méthode du gradient se ramène au calcul de l'approximation suivante avec la formule

$$u_j = \tilde{u}_j - \varkappa \sum_{i=j}^{N-1} \frac{\partial I_i}{\partial u_j} = \tilde{u}_j - \varkappa G_j, \tag{6.10}$$

où $\tilde{u}_j$ est l'approximation précédente, $\varkappa > 0$ le pas de la méthode.

On rappelle que $\partial I_i/\partial u_j$ sont les dérivées d'une fonction scalaire par rapport à un vecteur, c'est-à-dire sont des vecteurs de composantes $\frac{\partial I_i}{\partial u_j^1}, \ldots, \frac{\partial I_i}{\partial u_j^m}$.

Le changement (6.10) transforme la fonctionnelle (6.9) en une fonction de la variable scalaire $\varkappa$: $J = J(\varkappa)$. Le pas $\varkappa$ peut être choisi de manière à minimiser $J(\varkappa)$. Cette variante de la méthode du gradient s'appelle méthode de plus grande pente.

Calculons les vecteurs $G_i$. Remarquons à cet effet que la fonction $J(\tilde{u} + v)$ peut être mise sous la forme

$$J(\tilde{u} + v) = J(\tilde{u}) + \delta J + O(v^2),$$

où

$$\delta J = \sum_{i=0}^{N-1} (G_i, v_i),$$

$v_i$ étant l'accroissement du vecteur $u_i$. Posons $x_i = \tilde{x}_i + y_i$, $u_i = \tilde{u}_i + v_i$ dans les équations (6.3). Si l'on ne garde que les termes linéaires, on obtient

$$y_{i+1} = A_{i+1} y_i + B_{i+1} v_i, \tag{6.11}$$

où $A_{i+1}$, $B_{i+1}$ sont des matrices: $A_{i+1} = E + \tau \frac{\partial f(\tilde{x}_i, \tilde{u}_i)}{\partial x}$, $B_{i+1} = \tau \frac{\partial f(\tilde{x}_i, \tilde{u}_i)}{\partial u}$, $E$ est la matrice unité.

Eliminons successivement les variables de phase des équations (6.11). L'extrémité gauche de la trajectoire étant fixée, on a $y_0 = 0$

et par suite

$$
\begin{aligned}
y_1 &= B_1 v_0 = D_{1,0} v_0, \\
y_2 &= A_2 B_1 v_0 + B_2 v_1 = D_{2,0} v_0 + D_{2,1} v_1, \\
&\ldots\ldots\ldots\ldots\ldots\ldots \\
y_s &= D_{s,0} v_0 + D_{s,1} v_1 + \ldots + D_{s,s-1} v_{s-1},
\end{aligned}
\tag{6.12}
$$

où $D_{s,i}$ sont des matrices

$$
\begin{aligned}
D_{s,0} &= A_s A_{s-1} \ldots A_2 B_1, \\
D_{s,1} &= A_s A_{s-1} \ldots A_3 B_2, \\
&\ldots\ldots\ldots\ldots \\
D_{s,s-1} &= B_s.
\end{aligned}
$$

Transformons de façon analogue l'expression de la fonction $\delta J$:

$$\delta J(y, v) = \sum_{i=1}^{N-1} (d_i, y_i) + \sum_{i=0}^{N-1} (g_i, v_i), \tag{6.13}$$

où

$$d_i = \tau \frac{\partial F(\tilde{x}_i, \tilde{u}_i)}{\partial x}, \quad g_i = \tau \frac{\partial F(\tilde{x}_i, \tilde{u}_i)}{\partial u}.$$

Portons les expressions (6.12) de $y_s$ dans (6.13). Par des transformations évidentes, on obtient en définitive

$$\delta J = \sum_{i=0}^{N-1} (G_i, v_i), \tag{6.14}$$

où les vecteurs $G_i$ sont définis par

$$G_0 = \sum_{s=1}^{N-1} D^*_{s,0} d_s + g_0, \quad G_1 = \sum_{s=2}^{N-1} D^*_{s,1} d_s + g_1, \ldots, \tag{6.15}$$

et $D^*_{s,i}$ est la matrice transposée. Les dérivées de la fonctionnelle se définissent donc explicitement par les formules (6.14), (6.15).

D'après la méthode du gradient, $v_i = -\varkappa G_i$ ; donc

$$\delta J = -\varkappa \sum_{i=0}^{N-1} (G_i, G_i).$$

On peut construire par analogie d'autres schémas d'optimisation finidimensionnelle.

c) *Schémas possibles de programmation dynamique.* L'opération élémentaire nous a permis de mettre la fonctionnelle $J$ sous la forme (6.8). De telles fonctions sont dites *additives*: elles se représentent par une somme de termes dépendant chacun seulement de deux variables à indices successifs. Cette structure permet d'utiliser divers

schémas d'analyse séquentielle des variantes. Voyons l'un de ces schémas.

Nous avons déjà signalé que l'opération élémentaire permet de formuler le problème de commande optimale en termes de graphes: le problème se ramène en effet à la détermination du plus court chemin d'un graphe de forme spéciale. Passons aux détails.

Traçons les hyperplans $\Sigma_i$: $t = i\tau$, $i = 0, 1, 2, \ldots, N$, de l'espace $(x, t)$. Supposons que $x_i \in \Sigma_i$ et munissons chaque hyperplan d'un réseau de nœuds $x_i^s$, où $s$ est le numéro du nœud de l'hyperplan $\Sigma_i$. En reliant les nœuds $x_i^s$, $i = 0, 1, \ldots, N$; $s = 1, \ldots, M$ à l'aide de l'opération élémentaire, on obtient un graphe (fig. 2.2).

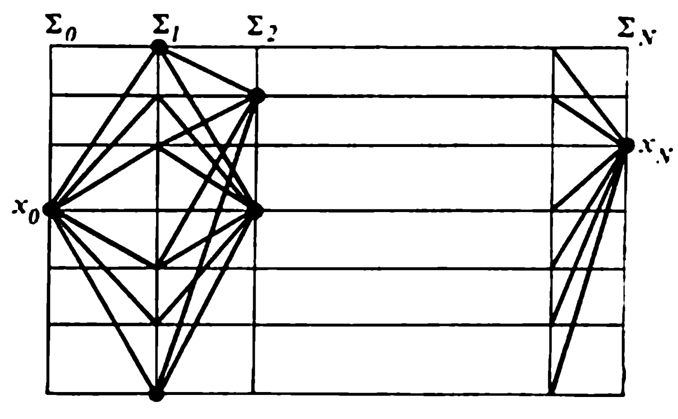



Fig. 2.2

Le problème est de trouver la plus courte ligne polygonale reliant les points $x_0$ et $x_N$ de ce graphe. On rappelle que par longueur d'un segment de ligne polygonale joignant les points $x_i^s$ et $x_{i+1}^k$ on a convenu d'entendre la quantité

$$\varphi_i(x_i^s,\ x_{i+1}^k) = l_{ks}(i) = \int_{t_i}^{t_{i+1}} F(\gamma_{i,\ i+1},\ u_{ks}(i))\, dt,$$

où $u_{ks}(i)$ est une commande construite à l'aide de l'opération élémentaire sur les nœuds donnés $x_i^s$, $x_{i+1}^k$.

REMARQUE. L'opération élémentaire définit la commande sur la portion de courbe joignant les points $x_i^s$, $x_{i+1}^k$. Mais nous avons déjà vu que dans la plupart des cas la commande est choisie approximativement. La résolution du problème de Cauchy

$$x(t_i) = x_i^s, \qquad \dot{x} = f(x,\ u_{ks}(i))$$

avec la commande choisie à l'aide de l'opération élémentaire nous donne une trajectoire qui ne passe pas exactement par le point $x_{i+1}^k$. Donc, les trajectoires

définies par les commandes $u = u\,(x_i^s,\ x_{i+1}^k)$ ne feront généralement qu'approcher la trajectoire passant par ces nœuds.

Exposons l'algorithme de recherche du plus court chemin sur le graphe spécial de la figure 2.2. Cet algorithme qui s'appelle «balai de Kiev» a été proposé par V. Mikhalévitch et N. Chor (cf. [53]). Considérons des points $x_1^{k_1}$ de l'hyperplan $\Sigma_1$ et désignons par $l_{k_10}(1)$ la distance de chacun de ces points à $x_0$. Considérons maintenant des nœuds $x_2^{k_2}$ de l'hyperplan $\Sigma_2$. La distance de chacun de ces points à un nœud de $\Sigma_1$ sera notée $l_{k_2k_1}(2)$. Quant à la longueur du chemin joignant les nœuds $x_2^{k_2}$ au point initial $x_0$ en passant par $x_1^{k_1}$, elle est égale à la somme

$$l_{k_10}(1) + l_{k_2k_1}(2).$$

Définissons maintenant le plus court chemin $l_{k_2}(2)$ joignant le point initial $x_0$ au point $x_2^{k_2}$. Il est égident que

$$l_{k_2}(2) = \min_{k_1}\,(l_{k_10}(1) + l_{k_1k_2}(2)) \qquad (6.16)$$

La relation (6.16) associe à chaque nœud de $\Sigma_2$ le plus court chemin reliant ce nœud au point initial $x_0$. Seuls ces chemins devront être retenus pour une analyse ultérieure, les autres ne présentent pas d'intérêt, car ils n'appartiennent pas au plus court chemin joignant les points $x_0$ et $x_N$. En effet, en vertu de la formule (6.8) la longueur totale du chemin joignant les extrémités de la trajectoire peut être représentée par la somme

$$l_{k_10}(1) + l_{k_2k_1}(2) + \ldots\ l_{k_{s+1}k_s}(s+1) + \ \ldots\ + l_{k_Nk_{N-1}}(N),$$

et celle du plus court chemin par la formule

$$J = \min_{k_{N-1}}\,(\ldots(\min_{k_2}\,(\min_{k_1}\,(l_{k_10}(1) + l_{k_2k_1}(2)) + \\ + l_{k_3k_2}(3))\ \ldots) + l_{k_Nk_{N-1}}(N)).$$

Donc, seule la ligne de longueur $l_{k_2}(2)$ peut être incluse dans le plus court chemin reliant les points $x_0$ et $x_N$.

La suite de l'algorithme est évidente. Après avoir déterminé les lignes de longueur $l_{k_2}(2)$, on cherche à l'étape suivante les plus courts chemins qui relient les nœuds $x_3^{k_3}$ au point initial. On obtient

$$l_{k_3}(3) = \min_{k_2}\,(l_{k_2}(2) + l_{k_3k_2}(3)), \qquad (6.17)$$

et ainsi de suite.

Cet algorithme demande un important temps machine, car il est lié à une longue opération de tri. Estimons sa taille. Désignons par $M_i$ le nombre de nœuds du plan $\Sigma_i$. Donc, à chaque pas $i+1$ on procède à un tri des variantes de l'ensemble $M_i$ de chemins passant

par le nœud fixe donné $x_{i+1}^{k_{i+1}}$, on choisit une variante à l'aide d'une formule (6.17), puis on la mémorise. Ainsi à chaque pas $i+1$, on doit mémoriser $M_{i+1}$ nombres $l_{k_{i+1}}(i+1)$. Pour déterminer les quantités $l_{k_{i+1}}(i+1)$ il faut calculer $M_i$ fonctions $l_{k_{i+1}k_i}(i+1)$, les ajouter à la quantité mémorisée $l_{k_i}(i)$ et comparer les résultats obtenus.

Supposons que cette procédure nécessite $M_i r$ opérations de machine. Donc, le nombre total $Q$ d'opérations de machine nécessaires à la réalisation de l'algorithme est estimé par la formule

$$Q = \sum_{i=0}^{N-1} M_i M_{i+1} r \leqslant M^2 r N, \quad M = \max_i M_i. \tag{6.18}$$

La quantité $M_i$ dépend de la dimension $n$ du vecteur $x$. Si l'on désigne par $p_i$ le nombre de nœuds sur chaque coordonnée, alors $M_i = p_i^n$. La majoration (6.18) devient alors

$$Q \sim p^{2n} r N, \quad p = \max_i p_i. \tag{6.19}$$

L'estimation (6.19) montre que la taille de cette variante de la méthode de programmation dynamique est peu sensible à l'accroissement de l'intervalle temporel (à l'accroissement du nombre $N$ dont elle dépend linéairement). En revanche, elle dépend fortement de la dimension du vecteur $x$.

Le volume des calculs nécessaires à la réalisation de ce schéma a suscité la création de nombreuses variantes simplifiées (méthode du tube errant, méthode des variations locales, etc.) qui font largement recette (cf. à ce propos [6]).

Les méthodes de cette nature sont très efficaces pour la résolution de nombreux problèmes de planification. Elles sont universelles en ce sens qu'on peut les utiliser sous les contraintes les plus générales lorsque les autres méthodes manquent de souffle.

## § 7. Problèmes de synthèse

Du schéma général de l'optimisation en deux étapes il appert que la détermination du programme optimal n'est que le premier pas de l'élaboration du système de commande. L'étape suivante est la construction d'un mécanisme de commande qui réaliserait la trajectoire programmée trouvée, c'est-à-dire qui assurerait la réalisation de l'objectif de la commande avec la plus grande précision, compte tenu des ressources disponibles.

Comme indiqué au § 3 de ce chapitre, ce problème se ramène au suivant : déterminer une commande correctrice $v$ telle que

$$I(v) \Rightarrow \min_v \tag{7.1}$$

sous la condition

$$\dot{x} = Ax + Bv + \xi, \tag{7.2}$$

$$x(0) = 0, \tag{7.3}$$

où $\xi = \xi(t)$ sont les perturbations extérieures. On admet que $\xi(t)$ est un processus aléatoire centré (c'est-à-dire que $\overline{\xi(t)} = 0$ pour tout $t$) dont on connaît entièrement la description stochastique.

REMARQUE. Dans le cas général, la condition (7.3) peut être remplacée par la condition plus générale $x(0) = \alpha$, où $\alpha$ est une variable aléatoire centrée.

La fonctionnelle $I(v)$ définit la précision de réalisation de l'objectif, par exemple la valeur de la variance.

Nous avons vu au § 3 que dans la recherche de la commande correctrice optimale, les fonctions $v$ ne doivent pas dépendre uniquement du temps, mais aussi de l'état du système, c'est-à-dire que

$$v = v(x, t). \tag{7.4}$$

De plus la commande correctrice est généralement soumise à des contraintes que nous écrirons sous la forme

$$v \in G. \tag{7.5}$$

Il n'existe pas de méthodes régulières pour la résolution de ce problème dans le cas général. Ce problème est bien plus compliqué qu'un problème ordinaire de commande optimale et l'on ne dispose non plus d'aucun résultat général, par exemple de conditions nécessaires de type principe du maximum, qui nous serve de point de départ à l'élaboration de méthodes de calcul efficaces.

REMARQUE. Cette assertion n'est pas tout à fait exacte. Pour la fonctionnelle (7.1) et la commande (7.4) on peut toujours former l'équation de Bellman et construire des procédures de résolution par les méthods de programmation dynamique. Mais la réalisation sur ordinateur de ces procédures n'est possible que dans des cas exceptionnels.

Le problème (7.1) à (7.5) s'appelle *problème de synthèse de la commande* ou *problème de détermination de l'opérateur de rétroaction*. Par ce terme on désigne la relation (7.4), car elle met en correspondance la quantité $x$ caractérisant l'écart de la trajectoire réelle par rapport à la trajectoire programmée (le programme optimal) et la commande correctrice $v$. La construction de l'opérateur de rétroaction est le problème clé de tout processus de commande.

On peut se tromper dans le choix du programme optimal. Ceci modifiera la valeur de la fonctionnelle: par exemple, la réalisation de l'objectif nous reviendra plus cher. Mais si l'opérateur de rétroaction est mal construit, c'est le processus de commande tout entier qui est compromis: le système perd de sa stabilité et risque de ne pas atteindre l'objectif fixé. Supposons qu'on ait programmé la tra-

jectoire de vol d'une fusée et qu'on ait trouvé le procédé le moins coûteux pour la mettre sur orbite. Cela ne résout pas le problème si le pilote automatique ne garantit pas la stabilité du mouvement de la fusée et si de petites perturbations aléatoires sont susceptibles de la renverser. Ceci concerne également les systèmes économiques et sociaux: la synthèse de l'opérateur de commande (l'élaboration des mécanismes de commande) est un problème majeur de l'analyse et de la construction des systèmes.

REMARQUE. Malgré l'importance de la construction de l'opérateur de rétroaction dans la théorie de la commande et le nombre de recherches consacrées à ce problème, aucune théorie générale de synthèse n'a été édifiée à ce jour. Ce problème est excessivement compliqué et seuls les cas élémentaires de systèmes linéaires ont été plus ou moins complètement étudiés.

Dans ce paragraphe nous passerons brièvement en revue quelques méthodes numériques de construction de l'opérateur de rétroaction, c'est-à-dire de construction de la fonction $v(x, t)$, et indiquerons un procédé d'estimation de la fonctionnelle $I$ qui, indépendamment de sa signification physique, sera appelée mesure de la précision de réalisation de l'objectif de la commande.

a) *Synthèse linéaire d'une rétroaction à coefficients constants.* L'absence de méthodes rigoureuses de résolution du problème de synthèse optimale (7.1) à (7.5) a contraint les chercheurs à concentrer l'essentiel de leurs efforts sur le développement de méthodes de réduction de ce problème à des problèmes plus simples permettant de déterminer les caractéristiques du mécanisme de rétroaction qui garantirait des « résultats satisfaisants ».

Ces méthodes se ramènent en principe à la restriction de la classe des fonctions admissibles $v(x, t)$, c'est-à-dire à la construction de collections de fonctions $v(x, t)$ fournissant la solution du problème de synthèse. L'opérateur de rétroaction construit à l'aide de ces fonctions ne sera plus optimal au sens du problème initial (7.1) à (7.5). Aussi, dans de tels cas parlerons-nous non pas de synthèse optimale, mais de synthèse possible (ou admissible, ou virtuelle).

Nous avons déjà introduit l'un de ces procédés au début de ce paragraphe lorsque nous avons remplacé la fonctionnelle (7.1) par la condition de stabilité et avons construit l'opérateur de rétroaction sous forme d'une fonction linéaire des coordonnées de phase:

$$v = Lx, \tag{7.6}$$

où $L = (l_{ij})$ est une matrice dont les éléments sont des constantes appelées coefficients d'amplification.

La solution de ce problème n'est pas unique et ne fournit qu'un système de contraintes que doivent vérifier les coefficients d'amplification $l_{ij}$, par exemple un système de doubles inégalités de la forme

$$l_{ij}^{-} \leqslant l_{ij} \leqslant l_{ij}^{+}. \tag{7.7}$$

Cette non-univocité fait le jeu des ingénieurs, car elle leur permet de faire varier dans des limites données les valeurs des paramètres de construction et de satisfaire les diverses conditions qui n'ont pas été formalisées ou incluses dans la position initiale du problème.

Mais la réduction à un problème de stabilité n'a de raison d'être que lorsque le temps $T$ de fonctionnement du système est assez long.

La suite naturelle de ces idées lors de la construction des mécanismes de commande sur un intervalle de temps fini est la réduction des problèmes de synthèse à des problèmes de programmation non linéaire. Elucidons le schéma de cette réduction pour le cas où la fonctionnelle $J$ a la forme de l'espérance mathématique

$$J = \overline{(x(T),\ x(T))}. \tag{7.8}$$

Cherchons la solution du problème de rétroaction sous la forme (7.6), où $L = (l_{ij})$ est une matrice constante (comme dans le problème de l'opérateur de rétroaction réalisant un mouvement stable du système).

Désignons par $G(t, \tau) = (g_{ij}(t, \tau))$ la matrice de Green de l'équation $\dot{x} = (A + BL)x$. Les conditions initiales étant supposées nulles, la solution de l'équation

$$\dot{x} = (A + BL)x + \xi$$

peut être mise sous la forme

$$x(t) = \int_0^t G(t, \tau)\,\xi(\tau)\,d\tau. \tag{7.9}$$

La fonctionnelle (7.8) devient maintenant:

$$J = \int_0^T \int_0^T G(t, \tau)\,G(t, s)\,\overline{\xi(\tau)\,\xi(s)}\,d\tau\,ds. \tag{7.10}$$

Le processus aléatoire $\xi(t)$ est supposé connu, donc il en sera de même de tous les éléments de la matrice de corrélation (ou des covariances) $K = (k_{ij}(\tau, s))$, où $k_{ij} = \overline{\xi^i(\tau)\,\xi^j(s)}$. S'agissant des éléments de la matrice $G(t, \tau)$, ils dépendent des coefficients d'amplification $l_{ij}$. En se donnant la matrice $L$, on peut grâce aux formules (7.9), (7.10), calculer la valeur de la fonctionnelle. Mais il faut auparavant déterminer la matrice de Green, c'est-à-dire résoudre $n$ problèmes de Cauchy ($n$ étant la dimension du vecteur $x$).

Donc, la fonctionnelle $J$ est une fonction des coefficients d'amplification $l_{ij}$, une fonction qui est définie par l'intermédiaire de la solution d'un système d'équations différentielles linéaires. Mais comme la variable de phase $x(t)$ dépendra de façon non linéaire des

coefficients d'amplification $l_{ij}$, la résolution du problème de synthèse se ramène à celle d'un problème de programmation non linéaire. A partir de la fin des années soixante cette approche a fait largement recette, en dépit du fait que la fonctionnelle $J$ ne soit pas généralement convexe. La résolution du problème est souvent menée à bien par cette méthode (cf. [58, 59]), car dans les problèmes de synthèse la précision n'est généralement pas très élevée.

Le schéma de résolution proposé s'appuie sur deux hypothèses. *Primo*, on suppose que les coefficients d'amplification sont constants. Cette hypothèse arrange l'ingénieur projetant le système de commande, puisqu'elle lui permet d'utiliser un schéma relativement simple. Mais le problème de savoir dans quelle mesure ce système permet d'assurer la précision de réalisation d'une commande proche de l'optimale est un problème dont la solution ne peut être acquise qu'empiriquement, par exemple par des calculs sur ordinateur.

*Secundo*, on admet que l'opérateur de rétroaction dépend linéairement de l'écart du vecteur de phase par rapport à la trajectoire programmée. Cette hypothèse doit aussi être justifiée.

Des tentatives de rejet de ces hypothèses ont été faites dans de nombreux travaux.

b) *Synthèse linéaire optimale.* Considérons de nouveau le système linéaire

$$\dot{x} = Ax + v + \xi. \tag{7.11}$$

Cherchons la commande dans la classe des fonctions linéaires $v = Lx$ en admettant que les éléments de la matrice $L = (l_{ij})$ sont des fonctions du temps. Imposons à ces éléments les contraintes suivantes:

$$l_{ij}(t) \in D_{ij}, \tag{7.12}$$

où $D_{ij}$ sont des ensembles. Dans les problèmes d'ingénieurs ces contraintes sont généralement de la forme (7.7).

Admettons que le processus aléatoire $\xi(t)$ est donné par son expression canonique:

$$\xi(t) = \sum_{i=1}^{m} c_i \varphi_i(t), \tag{7.13}$$

où $\varphi_i(t)$ est un système donné de fonctions vectorielles de dimension $n$, $c_i$ des variables aléatoires indépendantes scalaires dont les caractéristiques statistiques sont connues, et en outre $\overline{c_i} = 0$ (le processus aléatoire est centré).

Cherchons le vecteur $x(t)$ sous la forme d'une somme

$$x(t) = \sum_{i=1}^{m} c_i \chi_i(t), \tag{7.14}$$

où $\chi_i(t)$ sont des fonctions vectorielles inconnues.

En portant les expressions (7.13) et (7.14) dans l'équation initiale (7.11), on obtient

$$\sum_{i=1}^{m} c_i\,[\dot{\chi}_i-(A+L)\,\chi_i-\varphi_i]=0. \qquad (7.15)$$

Pour que cette égalité soit vraie quels que soient les réalisations de $c_i$ et $t \in [0,\ T]$, il est nécessaire et suffisant que les crochets soient nuls, c'est-à-dire que les fonctions $\chi_i(t)$ soient solutions du $n \times m$-système d'équations différentielles ordinaires

$$\dot{\chi}_i=(A+L)\,\chi_i+\varphi_i, \qquad i=1,\ 2,\ \ldots,\ m. \qquad (7.16)$$

Comme $x(0)=0$, les fonctions vectorielles $\chi_i(t)$ doivent également satisfaire des conditions initiales nulles:

$$\chi_i(0)=0.$$

Considérons maintenant l'expression de la fonctionnelle (7.8). En y portant (7.14), on aura

$$J=\sum_{j=1}^{n}\overline{(x^j)^2(T)}=\sum_{j=1}^{n}\overline{\Big(\sum_{i=1}^{m}c_i\chi_i^j\Big)^2}\,. \qquad (7.17)$$

Vu que $c_i$ sont des variables aléatoires indépendantes, c'est-à-dire que $\overline{c_ic_k}=0$ pour $i \neq k$, l'expression (7.17) devient

$$J=\sum_{j=1}^{n}\sum_{i=1}^{m}\overline{c_i^2}\,(\chi_i^j)^2. \qquad (7.18)$$

Etant donné que $c_i$ sont des variables aléatoires dont la loi de probabilité est connue, les $\overline{c_i^2}$ sont des nombres donnés et la fonctionnelle (7.18) est une fonction déterministe des valeurs finies $\chi_i^j(T)$ des fonctions inconnues. Donc, le problème de synthèse linéaire optimale avec les contraintes (7.12) a été ramené à un problème de commande optimale qui consiste à déterminer les fonctions $l_{ij}(t)$ minimisant la fonctionnelle (7.18) sous les conditions (7.16), (7.12).

La résolution de ce problème nous donne les coefficients d'amplification en fonction du temps. Ce problème est certes de dimension assez élevée, mais étant à une extrémité libre, il est justiciable de méthodes efficaces du type méthode de Krylov-Tchernoousko.

Malheureusement, ce procédé de réalisation de l'opérateur de rétroaction ne peut être utilisé que pour des contraintes de la forme (7.12). Dans la pratique, on a plus souvent affaire à des contraintes d'une autre forme, par exemple

$$|\,v^i\,| \leqslant a_i. \qquad (7.19)$$

On pourrait certes tenter de développer des méthodes efficaces

de résolution de tels problèmes par l'introduction des fonctions de pénalisation. Mais la méthode des fonctions de pénalisation, qui est largement répandue et bien élaborée pour les problèmes déterministes de théorie des programmes optimaux, n'est pas encore bien au point pour les problèmes de synthèse et sa discussion a commencé il y a peu de temps.

c) *Synthèse dans les problèmes à fonctionnelle quadratique.* Jusqu'à maintenant nous avons traité des problèmes relatifs à la construction de l'opérateur de rétroaction sous forme d'une fonction linéaire des variables de phase. L'hypothèse de linéarité restreint les possibilités de la commande; les opérateurs de rétroaction qui utilisent des dépendances fonctionnelles plus compliquées peuvent réaliser généralement l'objectif de la commande avec une plus grande précision que le meilleur des opérateurs linéaires. Une question vient immédiatement à l'esprit: existe-t-il des systèmes pour lesquels la solution du problème de synthèse optimale est un opérateur de rétroaction linéaire? Comment sont ces systèmes? La réponse est fournie par le théorème suivant *).

THÉORÈME. *Soit donné le système linéaire* (7.11) *sans contraintes sur la commande. Dans ces conditions la fonctionnelle quadratique est minimisée par une fonction linéaire* (7.6), *où les éléments $l_{ij}$ sont des fonctions du temps.*

Précisons maintenant l'expression de la fonctionnelle. Jusqu'ici nous avons étudié des fonctionnelles quadratiques de la forme

$$J = \overline{(x(T),\ x(T))} \tag{7.20}$$

ou

$$J = (\overline{x(T),\ R(x(T))}) \tag{7.20'}$$

pour estimer la précision de réalisation de l'objectif de la commande. Mais la minimisation de telles fonctionnelles en l'absence de contraintes sur la commande est triviale.

En effet, si l'on admet que la commande n'est soumise à aucune contrainte, on peut pendant un certain temps ne pas gouverner le système. On peut attendre que des perturbations extérieures $\xi(t)$ le fassent dévier du régime programmé (de la trajectoire programmée qui dans le cas présent correspond à l'origine des coordonnées $x \equiv 0$) et ensuite appliquer une impulsion pour compenser cette déviation. Plus on tardera à appliquer cette impulsion (c'est-à-dire plus on s'approchera de la fin du processus $t = T$) et plus sera grande la précision d'accès à l'objectif de la commande, c'est-à-dire au point $x_T \equiv 0$.

Sous cette forme, le résultat acquis présente peu d'intérêt, dans

*) Voir démonstration dans [6]

la mesure où en réalité les impulsions correctrices sont toujours bornées et il est impossible de réaliser une impulsion supérieure à une quantité donnée. Mais nous ne savons pas encore résoudre le problème de synthèse avec des contraintes de type inégalités sur la commande sans le cas général. Quant à la théorie des fonctions de pénalisation, elle n'est pas encore en mesure d'aborder de tels problèmes. Aussi, dans les problèmes d'ingénieurs remplace-t-on généralement la fonctionnelle (7.20) par la fonctionnelle

$$J = \overline{(x(T),\ x(T))} + \int_0^T \lambda(t)\, \overline{(v(t),\ v(t))}\, dt, \tag{7.21}$$

où $\lambda(t) > 0$ est une fonction donnée.

A la fin du numéro précédent, nous avons suggéré d'appliquer les fonctions de pénalisation à la résolution des problèmes de synthèse avec des contraintes sur la commande. Ceci est-il valable dans le cas étudié?

Le terme supplémentaire de (7.21) ressemble à une fonction de pénalisation. Mais la ressemblance n'est qu'apparente. Si le facteur $\lambda(t)$ de l'expression de la fonction de pénalisation tend vers $\infty$, alors la solution du problème avec la fonction de pénalisation tend vers celle du problème avec une contrainte dont nous avons pénalisé la violation. Comme l'intégrale de (7.21) n'est pas liée à la forme de la contrainte, cette assertion n'est pas vraie ici. Néanmoins cette intégrale a un sens: elle limite un éventuel accroissement des commandes et donne des résultats acceptables. En effet en faisant varier $\lambda(t)$ on assure en principe la réalisation des contraintes sur les commandes.

Nous avons formulé un théorème valable pour toute fonctionnelle quadratique et notamment pour les fonctionnelles de la forme (7.21).

Appliquons les raisonnements qui nous ont conduits à ce résultat au cas de la fonctionnelle (7.21). Remplaçons l'équation initiale (7.11) par le système d'équations aux différences

$$x_{k+1} = \Phi_k x_k + w_k + f_k, \qquad k = 0, 1, \ldots, N-1, \tag{7.22}$$

où $N$ est le nombre de sous-intervalles partiels de l'intervalle $[0, T]$. Si l'on se sert d'un schéma aux différences élémentaire du premier ordre d'approximation, on obtient

$$\Phi_k = E + A(t_k)\,\tau, \qquad w_k = v(t_k)\,\tau, \qquad f_k = \xi(t_k)\,\tau,$$

où $\tau = T/N$ est le pas temporel, $E$, la matrice unité. Transformons la fonctionnelle (7.21) de façon analogue:

$$J = \overline{(x_N,\ x_N)} + \sum_{k=0}^{N-1} d_k \overline{(w_k,\ w_k)}. \tag{7.23}$$

Nous sommes ainsi conduits à un problème de minimisation de la fonctionnelle (7.23) sous les contraintes (7.22), c'est-à-dire que nous devons trouver des quantités $w_0, \ldots, w_{N-1}$ vérifiant les conditions (7.22) et minimisant la fonction (7.23). Résolvons ce problème par la méthode de programmation dynamique.

Supposons que le système se trouve dans l'état $x_{N-1}$, c'est-à-dire que le processus se trouve à un pas de sa fin. Posons le problème auxiliaire suivant : trouver une commande $w_{N-1}$ minimisant la fonctionnelle (7.23). Le système se trouve déjà dans l'état $x_{N-1}$ et comme nous ne pouvons plus modifier cet état, le seul moyen qui nous reste pour influencer le résultat final est de choisir (convenablement) la commande sur le dernier intervalle temporel partiel.

Posons

$$\hat{J}_N = (x_N, x_N) + \sum_{k=0}^{N-1} d_k(w_k, w_k) =$$
$$= (x_N, x_N) + d_{N-1}(w_{N-1}, w_{N-1}) + \sum_{n=0}^{N-2} d_k(w_k, w_k). \quad (7.24)$$

Le dernier terme de cette expression a déjà été choisi par hypothèse.

Portons dans (7.24) l'expression (7.22) de $x_N$. Nous obtenons

$$\hat{J}_N = (\Phi_{N-1}x_{N-1}, \Phi_{N-1}x_{N-1}) + 2(w_{N-1}, \Phi_{N-1}x_{N-1}) +$$
$$+ 2(f_{N-1}, \Phi_{N-1}x_{N-1}) + 2(w_{N-1}, f_{N-1}) + (w_{N-1}, w_{N-1}) +$$
$$+ (f_{N-1}, f_{N-1}) + d_{N-1}(w_{N-1}, w_{N-1}) + \sum_{n=0}^{N-2} d_k(w_k, w_k).$$

Calculons l'espérance mathématique conditionnelle $J_{N-1}$ de la quantité $\hat{J}_N$ en admettant que $x_{N-1}$ et $\sum_{n=0}^{N-2} d_k(w_k, w_k)$ sont fixes et en tenant compte de ce que $\overline{f_{N-1}} = 0$:

$$J_{N-1}(x_{N-1}) = (x_{N-1}, R_{N-1}x_{N-1}) + 2(w_{N-1}, \Phi_{N-1}x_{N-1}) + (w_{N-1}, w_{N-1}) +$$
$$+ \overline{(f_{N-1}, f_{N-1})} + d_{N-1}(w_{N-1}, w_{N-1}) + \sum_{n=0}^{N-2} d_k(w_k, w_k),$$

où

$$R_{N-1} = \Phi^*_{N-1}\Phi_{N-1}.$$

Déterminons le minimum de $J_{N-1}$ à partir de la condition

$$\frac{\partial J_{N-1}}{\partial w_{N-1}} = 0.$$

Nous pouvons nous servir de cette condition, puisque les commandes, c'est-à-dire les quantités $w$, ne sont astreintes à aucune contrainte. Cette condition nous donne une équation pour la détermination de

$w_{N-1}$:

$$\Phi_{N-1}x_{N-1} + w_{N-1} + d_{N-1}w_{N-1} = 0,$$

d'où

$$w_{N-1} = -\frac{\Phi_{N-1}x_{N-1}}{1+d_{N-1}} = B_{N-1}x_{N-1}. \qquad (7.25)$$

Ainsi, au dernier pas du processus nous devons considérer que la commande $w_{N-1}$ est une fonction linéaire de la variable de phase $x_{N-1}$. En portant (7.25) dans l'expression de $J_{N-1}$ on trouve

$$J_{N-1} = \frac{d_{N-1}}{1+d_{N-1}}(x_{N-1}, R_{N-1}x_{N-1}) + \overline{(f_{N-1}, f_{N-1})} + \sum_{n=0}^{N-2} d_k(w_k, w_k). \qquad (7.26)$$

Tirons maintenant quelques conclusions sur l'étape initiale de résolution du problème de synthèse par la méthode de programmation dynamique. Quel que soit l'état du système avant le dernier pas et indépendamment de son passé, la commande $w_{N-1}$, la meilleure de toutes les commandes admissibles à ce pas, se détermine à l'aide de la formule (7.25) et c'est une fonction linéaire de l'écart $x_{N-1}$ par rapport à la trajectoire programmée.

En déterminant la valeur optimale de $w_{N-1}$, on obtient simultanément les conditions qui doivent être remplies par les autres commandes $w_0, \ldots, w_{N-2}$. Celles-ci doivent minimiser une fonctionnelle $J_{N-1}$ définie par la formule (7.26). Dans cette fonctionnelle figurent un terme, le dernier, qui caractérise l'écart du vecteur de phase et une intégrale qui pénalise pour une valeur trop élevée de la commande. La fonctionnelle (7.26) se distingue de l'initiale par le fait que son dernier terme est défini maintenant non plus pour $k = N$, mais pour $k = N - 1$. En déterminant la commande $w_{N-1}$ sur le dernier intervalle, nous avons transféré la condition imposée à la fin du processus d'un pas vers l'origine, c'est-à-dire au point $t_k = N - 1$.

La minimisation de la fonctionnelle (7.26) est un problème qui ne diffère en rien de celui considéré et l'on peut lui appliquer les mêmes raisonnements. En d'autres termes, nous devons de nouveau admettre que l'état précédent, c'est-à-dire le vecteur $x_{N-2}$, est connu, porter dans (7.26) l'expression suivante de $x_{N-1}$:

$$x_{N-1} = \Phi_{N-2}x_{N-2} + w_{N-2} + f_{N-2},$$

transformer l'expression (7.26) et trouver le minimum d'une fonction quadratique en $w_{N-2}$. Ce qui nous donne

$$w_{N-2} = B_{N-2}x_{N-2}, \qquad (7.27)$$

c'est-à-dire qu'à l'avant-dernier pas du processus la commande est

encore une fonction linéaire de l'écart $x_{N-2}$, quel que soit le passé du processus.

Nous devons ensuite composer la fonctionnelle $J_{N-2}$ pour déterminer les commandes $w_0, w_1, \ldots, w_{N-3}$. Nous obtenons

$$J_{N-2}=(x_{N-2},\ R_{N-2}x_{N-2})+\frac{d_{N-1}}{1+d_{N-1}}\overline{(f_{N-2},\ R_{N-1}f_{N-2})}+$$
$$+\overline{(f_{N-1},\ f_{N-1})}+\frac{2d_{N-1}}{1+d_{N-1}}\overline{(f_{N-2},\ \Phi^{*}_{N-1}f_{N-1})}+\sum_{k=0}^{N-3} d_k\,(w_k,\ w_k),$$

où la matrice $R_{N-2}$ se calcule trivialement, etc. Cette procédure nous donne une suite de commandes de la forme

$$w_i = B_i x_i, \tag{7.28}$$

qui sont des fonctions linéaires des états correspondants. La formule (7.28) nous donne la solution complète du problème.

Les raisonnements développés ne sont pas une démonstration rigoureuse du théorème. En effet, nous avons prouvé le théorème pour le cas seulement d'équations aux différences finies. Le passage à la limite pour $\tau \to 0$ (ou $N \to \infty$) est loin d'être trivial. Et bien que l'optimalité de l'opérateur linéaire de rétroaction dans les problèmes à fonctionnelle quadratique est un fait connu probablement depuis la fin des années quarante, sa démonstration rigoureuse n'a pu être produite qu'au début des années soixante par R. Caleman.

Le résultat acquis est d'une grande importance pratique, pas seulement parce que nous pouvons expliciter l'opérateur de rétroaction pour cette classe de problèmes, mais parce que nous avons trouvé une classe de systèmes pour lesquels est notamment optimal un opérateur de rétroaction linéaire. Or l'écrasante majorité des mécanismes de commande sont construits justement selon ce type. Connaissant la classe de systèmes pour lesquels ces opérateurs sont optimaux, on peut estimer le degré de conformité des mécanismes créés au modèle optimal des systèmes commandés.

d) *Synthèse de l'opérateur de rétroaction par une prévision.* L'une des principales difficultés dans la construction des mécanismes de commande réside dans le fait qu'il faut tenir compte de la stochasticité des perturbations extérieures et opérer avec des équations stochastiques. L'absence d'outil mathématique efficace nous force à simplifier le problème (en le remplaçant par exemple par un problème linéaire), à restreindre les classes d'opérateurs parmi lesquels on cherche l'opérateur optimal, etc. Mais en dépit de toutes ces simplifications, on tente toujours en fin de compte de ramener le problème à une suite de problèmes d'optimisation déterministes.

Dans ce numéro on se propose de décrire une méthode de réalisation du principe de rétroaction en s'appuyant sur la possibilité de

trouver un programme optimal sans admettre la linéarité du système initial.

Considérons un système commandé de forme assez générale

$$\dot{x} = f(x, u, \xi, t), \tag{7.29}$$

où la commande est soumise à la contrainte

$$u \in U, \tag{7.30}$$

et la qualité de la commande est estimée par la fonctionnelle

$$J_t = F(x(T)) + \int_t^T \varphi(x, u(s), s)\, ds\, ^{*}), \tag{7.31}$$

où $F$ et $\varphi$ sont des fonctions scalaires données.

On admettra pour fixer les idées que l'état initial du système est fixé :

$$x(0) = x_0. \tag{7.32}$$

Partageons l'intervalle $[0, T]$ en $N$ sous-intervalles de longueur $\tau$: $\tau_0 = 0$, $\tau_N = T$, $\tau_i = i\tau$. Construisons la prévision $\xi^*(t)$ du processus aléatoire $\xi(t)$ á l'instant $\tau_i$ sur le vu des observations des perturbations extérieures sur l'intervalle $[0, \tau_i]$. Cette prévision est égale à l'espérance mathématique conditionnelle de la variable $\xi(t)$ sachant ses valeurs pour $t \leqslant \tau_i$:

$$\xi^*(t) = (\overline{\xi(t)}/\xi(s),\ s \leqslant \tau_i), \qquad t > \tau_i. \tag{7.33}$$

La détermination de $\xi^*(t)$ est en principe possible, puisque l'on a admis que la description probabiliste du processus $\xi(t)$ est entièrement connue. Mais dans nombre de cas concrets, par exemple lors de l'estimation des caractéristiques météorologiques, les calculs effectués par la formule (7.33) peuvent être très laborieux. Il faut alors utiliser une prévision d'expert ; en tout état de cause nous pouvons toujours admettre que nous disposons d'un opérateur associant aux observations de la fonction $\xi(s)$ sur l'intervalle $0 < s \leqslant \tau_i$ une fonction $\xi^*(t)$ pour $t > \tau_i$:

$$\xi^*(t) = \Pi_i \xi(s). \tag{7.34}$$

En portant la fonction $\xi^*(t)$ dans l'équation (7.29), on obtient le système déterministe

$$\dot{x} = f(x, u, \xi^*(t), t), \qquad x(\tau_i) = x_i. \tag{7.35}$$

*) L'expression (7.31) estime la commande sous réserve que soit connu l'état $x(t)$ à l'instant $t$. Nous verrons plus loin comment estimer le processus « globalement ».

Construisons la commande optimale $u_i(t)$ dans le problème (7.35) avec la fonctionnelle $J_{\tau_i}$. La fonction $u_i(t)$ dépendra visiblement de $x_i$, $\tau_i$, $\xi^*(t)$, soit

$$u_i = u_i(t, \tau_i, x_i, \xi^*(t)), \qquad t \geqslant \tau_i.$$

La commande ainsi obtenue sera appelée *commande programmée par morceaux.*

Si l'on dispose des opérateurs de prévisions $\Pi_i$ et d'une méthode économique de construction de la commande programmée par morceaux, alors le schéma général de commande se présente comme suit. A l'instant initial on construit la prévision

$$\xi^*(t) = \Pi_0 \xi(0)$$

et on cherche la commande $u_0(t)$ qui minimise la fonctionnelle

$$J_0 = F(x(T)) + \int_0^T \varphi(x, u(s), s)\, ds.$$

Cette commande est utilisée sur l'intervalle de temps $[0, \tau_1]$. A l'instant $\tau_1$ on fait une nouvelle prévision en tenant compte de la réalisation de $\xi(t)$ observée:

$$\xi^*(t) = \Pi_1 \xi(s), \qquad s \in [0, \tau_1].$$

On construit ensuite une nouvelle commande programmée $u_1(t)$ à l'aide de cette prévision. La commande $u_1(t)$ est utilisée sur l'intervalle $[\tau_1, \tau_2]$ pour déterminer une nouvelle prévision et une nouvelle commande minimisant la fonctionnelle

$$J_{\tau_1} = F(x(T)) + \int_{\tau_1}^T \varphi(x, u(s), s)\, ds,$$

et ainsi de suite.

La méthode de commande exposée s'appelle parfois *méthode du plan glissant.* Il est évident qu'elle réalise l'opérateur de rétroaction à tout instant de correction. Elle permet de trouver la commande dépendant de l'état dans lequel se trouve l'objet sous l'action des perturbations extérieures.

Cependant ce schéma de commande diffère sur un point essentiel de ceux étudiés précédemment. Jusqu'ici nous avons constamment procédé à une optimisation en deux étapes et avons construit la commande qui devait ramener le système à la trajectoire programmée initialement. La méthode du plan glissant n'implique pas le retour du système à une trajectoire donnée. La rétroaction introduite dans le système par la méthode du plan glissant est orientée immédiatement vers un objectif: l'objectif de la commande. La procédure de construction de la trajectoire programmée réalise

elle-même l'opérateur de rétroaction : le problème de commande n'est pas décomposé en une procédure de calcul du programme et une procédure de synthèse.

La mise en œuvre de tels principes ouvre des perspectives intéressantes pour le développement de méthodes efficaces de synthèse de l'opérateur de rétroaction. Ces méthodes sont commodes dans de nombreux domaines de l'activité humaine: dans la gestion des systèmes d'irrigation, dans la planification économique, dans la gestion de ressources, etc. Mais pour en tirer le plus grand profit, il est nécessaire de comprendre sous quelles conditions elles sont optimales et comment elles sont reliées aux méthodes optimales de commande dans le cas général.

Signalons tout d'abord que puisque nous étudions un problème stochastique, la fonctionnelle estimant la qualité de la commande doit être formulée dans les mêmes termes. Considérons des fonctionnelles de la forme

$$J=\overline{J}_0=\overline{F(x(T))}+\int_0^T \overline{\varphi(x,u(s),s)}\,ds. \qquad (7.36)$$

Supposons que le processus aléatoire $\xi(t)$ est un bruit blanc, c'est-à-dire un processus gaussien d'espérance mathématique nulle et de matrice de corrélation $K(t,s)=E\delta(t-s)$, où $E$ est la matrice unité et $\delta(t-s)$ la fonction de Dirac. Supposons encore que le processus (7.29) est linéaire en les coordonnées de phase, la commande et la perturbation, par exemple qu'il est de la forme (7.11), et que la fonctionnelle (7.36) est quadratique, c'est-à-dire de la même forme que les fonctionnelles étudiées dans ce paragraphe. Ces hypothèses nous permettent de trouver explicitement la commande optimale et la valeur de la fonctionnelle (7.36) par la méthode du numéro précédent. Ce problème se résout sans peine aussi par la méthode des commandes programmées par morceaux. Grâce à cela la méthode des commandes programmées par morceaux peut être comparée avec une méthode exacte donnant la solution optimale.

Il se trouve que (cf. [40]):

1. Lors de la prévision de la perturbation à des instants discrets, l'écart entre les valeurs de la fonctionnelle données par la méthode de la commande programmée par morceaux d'un côté et par la synthèse stochastique optimale de l'autre est positif et de l'ordre de $\Delta$ ($\Delta$ étant le plus grand intervalle séparant les instants de correction).

2. Si la prévision des perturbations est constamment corrigée, alors les valeurs ci-dessus sont confondues.

Ainsi, dans le cas d'un bruit blanc et de systèmes linéaires, la méthode exposée de construction de l'opérateur de rétroaction est une méthode d'acquisition de la solution optimale. On ne dispose

d'aucun autre résultat mathématique sur les systèmes de forme plus générale. Il est probablement assez difficile d'en obtenir, puisque la théorie des équations stochastiques non linéaires est encore à l'état embryonnaire. Dans le cas général cette méthode peut être traitée comme une procédure euristique commode pour la construction de l'opérateur de rétroaction pour des systèmes non linéaires avec des contraintes générales sur la commande.

La méthode des commandes programmées par morceaux possède un vaste champ d'applications: on peut notamment s'en servir pour construire des graphiques de distribution qui sont largement utilisés dans la gestion de diverses unités économiques. Nous reviendrons sur cette question dans l'avant-dernier chapitre.

e) *Quelques commentaires.* Nous avons exposé quelques méthodes de synthèse des systèmes de commande, axées sur l'utilisation des ordinateurs. Nous constatons qu'elles revêtent en général un caractère euristique: pratiquement pas de résultats rigoureux et d'estimations rigoureuses. C'est pourquoi il serait plus logique de parler de synthèse admissible, ou possible, que de synthèse optimale. Mais nous avons dans le même temps exhibé des méthodes de construction des mécanismes de commande. Donc, nous avons la possibilité de les tester à l'aide de l'ordinateur. Cette expérience sur ordinateur basée sur les modèles mathématiques ouvre une page nouvelle dans le développement des méthodes d'étude des systèmes commandés. En effet, si l'on dispose d'un quelconque mécanisme de commande (un pilote automatique) réalisé sous la forme d'un algorithme, on peut suivre son fonctionnement grâce à l'expérience sur ordinateur. Si l'on connaît les critères, on peut apprécier son efficacité et décider de sa validité.

L'expérience sur ordinateur permet donc de mettre à contribution des méthodes non formelles pour l'analyse des systèmes de commande. Ceci est d'autant plus important qu'en pratique, lors de la construction de systèmes techniques de commande et notamment lors de la construction de systèmes de gestion d'unités économiques, toute position classique du problème, ne présentant aucune faille sur le plan mathématique, est une schématisation assez conventionnelle du processus réel étudié.

A la lumière de ce qui vient d'être dit, on ne saurait surestimer la portée des méthodes de synthèse approchée des systèmes de commande. Mais optant pour la construction des opérateurs de rétroaction par des procédures euristiques, nous chargeons l'expérience sur ordinateur de nombreuses difficultés. L'organisation de cette dernière devient un important problème qui sera discuté dès le prochain chapitre.

CHAPITRE III

# SYSTÈMES CYBERNÉTIQUES ET SIMULATION

## § 1. Sur le terme « analyse des systèmes »

Dans l'avant-propos on a défini l'« analyse des systèmes » et affirmé que c'était une sorte de synthèse des idées et principes de la théorie de la recherche opérationnelle et des méthodes de la théorie de la commande avec tout l'arsenal des ordinateurs actuels. On pourrait même ajouter que l'analyse des systèmes est l'étape contemporaine de développement de ces disciplines. Puis conformément à cette thèse on a exposé les idées et conceptions fondamentales de la théorie de la recherche opérationnelle et de la théorie de la commande. On se propose maintenant de reprendre, à partir de ces positions, la discussion de l'analyse des systèmes et l'étude de ses liens avec les autres disciplines. Ceci est d'autant plus nécessaire que l'on se sert de trois notions relatives aux systèmes: l'analyse des systèmes, la théorie des systèmes et l'approche systémique. On identifie souvent ces notions, d'où une certaine confusion. Vu que dans la suite nous aborderons les méthodes de l'analyse des systèmes, il nous faut d'emblée préciser les termes que nous entendons utiliser.

Le mot « système » et les termes qui lui sont rattachés sont largement répandus. Cela est dû au fait que la nécessité d'étudier des systèmes complexes *) se fait de plus en plus impérieusement sentir en raison de la sophistication des constructions techniques, des technologies et de toutes les infrastructures manipulées par les économistes, les gérants et les ingénieurs.

L'étude des objets biologiques et des problèmes écologiques qui prennent de l'ampleur d'une année à l'autre conduit le chercheur à des systèmes complexes.

L'« analyse des systèmes » a été bâtie pour répondre à ce besoin d'étudier les systèmes complexes. Il est donc naturel comme nous l'avons déjà signalé de la traiter comme le prolongement de la recherche opérationnelle et de la théorie de la commande dans la mesure où l'un des problèmes clés de l'analyse des systèmes est le problème de prise de décision.

*) La notion de « système » est de celles qu'il est difficile de définir soigneusement. Par système on entend souvent un ensemble d'éléments liés (par exemple, un système de masses attractives). Dans cet ouvrage nous ne tenterons pas de définir rigoureusement un système, nous satisfaisant de la notion intuitive que chaque lecteur s'en fait.

Dans les deux premiers chapitres nous avons abordé des problèmes qui relèvent traditionnellement de la recherche opérationnelle et de la théorie de la commande. Nous avons en particulier étudié le problème d'indétermination et notamment d'indétermination des objectifs, et avons avancé des hypothèses et exposé des méthodes aidant à sa résolution et précisant ces objectifs. Mais l'outil (de la théorie de la recherche opérationnelle) mentionné est souvent insuffisant.

L'indétermination de l'objectif rencontrée jusqu'à maintenant résidait dans la multiplicité des critères. Il était difficile de mesurer et de comparer les divers impératifs, difficile aussi de formaliser la notion d'« objectif » et d'unifier les indices. Il n'est pas non plus exclu qu'il soit impossible de fixer un objectif avec un tant soit peu de précision ou que l'objectif poursuivi soit irréel. (L'économie abonde de tels exemples.) Dans ce cas on se tourne vers l'analyse des systèmes.

Supposons que l'on étudie les perspectives de développement énergétique de la Sibérie occidentale. Comment définir les objectifs? Certes on peut toujours ainsi formuler les besoins: beaucoup de carburant à moindre coût, etc. Cependant il faut avancer des indices plus ou moins précis et des objectifs réalistes qui soient compatibles avec les besoins du pays et qui soient réalisables avec les ressources disponibles. De tels problèmes ne s'inscrivent plus dans le schéma classique de la recherche opérationnelle. En effet, le plus important dans ces problèmes est de formuler les objectifs du projet. L'objectif cesse d'être un facteur exogène comme en théorie de la recherche opérationnelle ou en théorie de la commande et devient un objet autonome d'étude.

Nous entrons maintenant dans le domaine de la recherche opérationnelle dont le but est de définir l'objectif d'une opération (par exemple le projet de développement d'une région). Que faut-il à l'analyste pour définir, pour « formuler correctement » les objectifs réalistes dont la réalisation doit être assurée par les constructions créées ou l'unité de production ? Il est évident qu'il faut tout d'abord s'imaginer le fonctionnement de la future construction, comparer ses capacités avec les ressources mises à la disposition du responsable. Ceci ne peut être réalisé que par des modèles physiques (des maquettes) ou mathématiques.

Si donc nous voulons utiliser les mathématiques, il nous faut commencer par décrire le système de modèles et édifier un appareil mathématique qui permette d'analyser le processus considéré, de voir les conséquences de nos décisions, d'estimer nos possibilités dans les diverses alternatives, et seulement après formuler les objectifs. La complexité des systèmes projetés et étudiés conduit à l'élaboration de techniques de recherche spéciales basées sur la simulation — qui consiste à reproduire sur ordinateur le fonctionnement du système

projeté ou étudié à l'aide de modèles mathématiques spécialement conçus.

REMARQUE. La dernière assertion ne déprécit nullement les méthodes classiques de l'analyse. Plus, l'utilisation de la simulation suppose nécessairement un traitement préalable du modèle. Ces problèmes feront l'objet des deux prochains chapitres.

Signalons que l'étude de la dynamique du processus en vue de sonder les perspectives et d'ébaucher les objectifs n'est qu'un aspect de l'analyse des systèmes, un aspect qui est sans nul doute le plus important, mais qui n'épuise pas toutes les questions auxquelles il est en mesure de répondre. Bref, ce n'est que le premier pas. Le problème suivant consiste à réaliser les objectifs fixés, c'est-à-dire à formuler une suite de décisions dont la mise en œuvre assure la réalisation de ces objectifs (définit les paramètres de l'engin construit ou du projet).

Parmi les problèmes soulevés par l'élaboration d'un projet, la combinaison des aspects structurels et des aspects fonctionnels occupe une importante place. L'une des plus grosses difficultés est soulevée par la projection des structures hiérarchiques. Tout système plus ou moins complexe est agencé suivant un principe hiérarchique du fait que le traitement centralisé de l'information et la prise de décision sont souvent rendus impossibles par la grande masse d'information à recueillir et à traiter, par les retards et les déformations qui s'ensuivent, etc. S'il est question de projets de systèmes techniques, la tâche du projeteur consiste avant toute chose à élaborer un schéma fonctionnel (susceptible d'être réalisé de plusieurs manières) et à définir des objectifs partiels.

La situation se complique singulièrement dans le cas de systèmes économiques dont le fonctionnement des divers éléments est tributaire des capacités de gestion des responsables. A l'opposé de la machine, l'homme est guidé par des intérêts et des objectifs personnels et il ne suffit pas seulement au projeteur de formuler les objectifs à l'intention des échelons inférieurs. Encore faut-il être sûr que ces objectifs seront réalisés, c'est-à-dire que les échelons inférieurs exécuteront les décisions des échelons supérieurs. A cet effet, il faut mettre en place un « mécanisme » spécial *), car il ne suffit pas d'une commande ni d'un ordre pour réaliser un objectif. D'où la nécessité d'une théorie spéciale appelée à développer les principes de hiérarchisation de la gestion et les méthodes de leur analyse. La théorie des systèmes hiérarchiques qui étudie certains aspects de ce problème constitue l'une des plus importantes branches de l'analyse des systèmes.

*) Dans l'économie socialiste, il existe un problème spécial de construction d'un mécanisme économique (khozrastchet, système de récompense, etc.) dont l'objectif est de réaliser les plans économiques.

L'analyse des systèmes est donc une discipline qui développe les méthodes d'élaboration de projets de systèmes techniques, économiques, écologiques complexes, d'infrastructures, etc. L'analyse des systèmes englobe la théorie de la recherche opérationnelle et la théorie de la commande avec tous leurs arsenaux.

Vu qu'aucune analyse de systèmes complexes n'est concevable sans les ordinateurs, quand on parle des méthodes de l'analyse des systèmes, on comprend généralement des procédures utilisant les ordinateurs.

Le terme « théorie des systèmes » a connu une large diffusion à l'instar de l'« analyse des systèmes ». Malgré son large usage, ce terme est interprété différemment. De même (comme déjà indiqué) on n'arrive pas à donner une définition assez rigoureuse du terme « système ».

L'apparition de la « théorie des systèmes » est généralement rattachée au nom du célèbre biologiste L. Bertalanffy (cf. [21]) qui organisa dans les années cinquante au Canada un centre d'études des systèmes et publia un grand nombre de travaux dans lesquels il s'appliqua à dégager le point commun à toutes les organisations assez complexes de la matière de nature tant biologique que sociale. Ces questions furent abordées par d'autres savants bien avant Bertalanffy. Les plus importantes recherches fondamentales sont à mettre à l'actif de notre compatriote A. Bogdanov qui dès le début du siècle commença à bâtir la théorie de l'organisation *). Dans ce travail, il introduit la notion d'organisation comme une notion liminale. La matière existe dans le temps et dans l'espace. Elle est toujours organisée et cette organisation est inconcevable sans son support matiérel. A. Bogdanov explique la construction de sa théorie par le fait qu'en dépit de la fantastique diversité des matériaux existant dans la nature, le nombre de formes architecturales et de formes d'organisation est relativement peu élevé. Il prouve ceci sur d'innombrables exemples de nature physique différente. Il étudie non seulement la statique, mais analyse les diverses particularités des mécanismes de sélection gouvernant l'évolution de l'organisation.

Plus tard, la théorie de l'organisation (synonyme de théorie des systèmes) attira l'attention d'éminents savants soviétiques: I. Schmalhausen [63], V. Beklémichev, etc., qui apportèrent d'importantes contributions originales à l'interprétation de la notion

---

*) Pour plus de détails voir [22]. La deuxième édition parut à Moscou en 1923, la dernière en 1929, après la mort de l'auteur qui fut l'organisateur et le premier directeur de l'Institut de transfusion sanguine. Nombre de recherches parmi les plus récentes soulignent la transcendance de certains points de la tectologie sur les idées de la cybernétique. La tectologie est marquée par la vision mécaniste de Bogdanov.

d'organisation et situèrent la portée de cette notion dans la conception générale de l'évolution du monde matériel.

REMARQUE. A mon sens, l'apport des chercheurs russes et soviétiques est à bien des égards décisif dans la formation de la théorie de l'organisation et les innombrables références aux seuls travaux de Bertalanffy et de ses adeptes ne sont pas toujours justifiées, puisque la plupart des idées développées actuellement sont d'une manière ou d'une autre reflétées dans les travaux de A. Bogdanov et des autres savants soviétiques.

Donc, à la différence de l'analyse des systèmes, discipline axée sur la résolution de problèmes pratiques, la théorie des systèmes tiendrait plutôt d'une science méthodologique.

A l'usage on ne fait souvent pas de distinction entre l'analyse des systèmes et la théorie des systèmes. D'après ce qui précède, ces deux termes ne doivent pas être confondus.

Mais l'analyse des systèmes et la théorie des systèmes sont encore loin de couvrir toute la terminologie systémique forgée ces dernières décennies.

Nous avons déjà souligné l'existence d'une notion: l'approche systémique qui est encore plus vague et plus imprécise. Ceci n'empêche qu'elle reflète des tendances qui se sont surtout manifestées dans les années d'après-guerre.

Dans le développement de toute science, se dégagent toujours nettement deux aspects: l'analyse et la synthèse. Nous constatons toujours d'une part une propension à l'analyse, à l'étude de faits concrets, à leur investigation en profondeur, à l'appréhension de la quintessence du phénomène considéré, etc., et de l'autre une tentative d'édifier des théories synthétiques permettant de rassembler des faits disparates, d'entrevoir les perspectives de développement des divers processus et leurs liens avec d'autres phénomènes, d'étudier leur interdépendance, etc.

La création des théories synthétiques s'accompagne parfois d'une perte d'information: les faits ne peuvent pas être tous incorporés dans un schéma unique, l'arsenal des méthodes utiles ne s'adapte pas toujours d'emblée au nouveau système de conceptions. Par exemple, bien que géocentrique, la théorie de Ptolémée n'en fournissait pas moins un procédé de calcul de la position des planètes sur la voute céleste. La théorie de Copernic était héliocentrique, mais les premiers temps elle n'indiqua aucune méthode de détermination de la position des planètes et par conséquent présentait peu d'intérêt pratique (par exemple pour la navigation maritime) contrairement à la théorie de Ptolémée. Ce n'est qu'après les travaux de Kepler qu'elle fut dotée des armes qui lui permirent de remplacer entièrement la théorie de Ptolémée.

Ces deux aspects connurent des fortunes différentes au long des siècles sans jamais cesser de se côtoyer. Le désir de dissocier les disciplines, de chercher de nouveaux faits, de faire d'un fait l'uni-

que objet de recherche scientifique est d'une grande importance. Dans le même temps, sont apparues de nouvelles branches connexes dans lesquelles il est impossible de démarquer une science de l'autre: la chimie de la physique ou de la biologie, etc.

Le rôle des constructions synthétiques a particulièrement grandi ces dernières décennies. Le besoin non seulement d'étudier un événement, mais d'établir ses liens avec d'autres faits a conduit à l'apparition du terme spécial d'approche systémique. Ce besoin est aujourd'hui si vif qu'on voit apparaître non seulement de nouvelles disciplines scientifiques à la jonction de certaines sciences naturelles, mais aussi des recherches qui relèvent dans une part égale des sciences naturelles et des sciences sociales. L'intérêt manifesté pour ces constructions synthétiques est lié aux possibilités croissantes du traitement de l'information. L'analyste a apparemment toujours cherché à aborder les faits étudiés, dans la mesure du possible, sous un angle systémique, mais il n'était pas toujours assez armé pour le faire. Avec l'avènement des ordinateurs, ses possibilités se sont fortement accrues. D'où par voie de conséquence cette tendance à étudier un événement en profondeur sans le détacher des autres événements.

Cette approche est constamment stimulée par les besoins de la pratique qui pose des projets de plus en plus compliqués impliquant l'analyse de problèmes interdisciplinaires.

REMARQUE. L'approche systémique, comme nous le verrons, est un principe méthodologique général. Son aspect gnoséologique est la théorie des systèmes, son instrument, l'analyse des systèmes. Cette division est assez conventionnelle et, comme nous le verrons plus bas, les questions des instruments ne peuvent (et ne doivent) pas et de loin être toujours différenciées des problèmes philosophiques.

Dans cet ouvrage nous ne toucherons pas aux problèmes généraux de méthodologie et porterons notre attention seulement sur l'analyse des systèmes, ses méthodes et sur ce qui la distingue essentiellement des autres disciplines, savoir son besoin d'unifier les méthodes d'analyse formelles et non formelles. Nous montrerons aussi comment les méthodes rigoureuses qui utilisent les ressources mathématiques modernes s'inscrivent naturellement dans le mode de pensée littéraire et scientifique.

## § 2. Problèmes de simulation

Les méthodes de l'analyse des systèmes (au même titre que les méthodes de la recherche opérationnelle et de la théorie de la commande) se basent sur la description mathématique des divers faits, événements, processus. Nos connaissances sont toujours relatives, donc toute description ne reflète que certains aspects des phénomènes et n'est jamais absolument complète.

Le terme « modèle » est passé dans le langage courant. La notion de « modèle » admet plusieurs interprétations, il existe une classification des modèles, etc. L'analyse détaillée de cette notion n'entre pas dans le cadre de cet ouvrage. Par les vocables « modèle », « description par un modèle », on comprendra une description qui reflète précisément les particularités du processus étudié qui intéressent l'analyste. La fidélité et la qualité de cette description sont définies en premier lieu par la conformité du modèle aux normes imposées à la recherche ainsi que par la concordance des résultats acquis à l'aide du modèle avec ceux obtenus par l'observation du processus réel.

Les modèles décrits avec le langage mathématique seront dits modèles mathématiques. Ce seront les seuls modèles envisagés dans la suite.

REMARQUE. Toute discipline scientifique est confrontée à des descriptions approximatives par des modèles. Mais ces modèles peuvent utiliser les langages (et symboles) les plus divers. Pour les distinguer des modèles mathématiques, on les qualifie de « modèle consistant », « modèle verbal », etc.

L'étude d'un modèle mathématique est toujours liée à une « algèbre », c'est-à-dire à des règles d'opérations sur les objets étudiés qui reflètent les relations entre les causes et les effets. Si cette algèbre est suffisamment développée, on dit qu'a été créée une théorie dans le cadre de ce modèle. Certains faits de théorie — les assertions, les théorèmes — sont parfois appelés lois (deuxième loi de Newton, loi de Stokes, etc.). Exactement de même, de nombreuses thèses à caractère phénoménologique, bien vérifiées par l'expérience, sont érigées en lois. C'est pourquoi on dit parfois qu'une théorie (ou un modèle) est fondée sur des lois.

La construction de modèles mathématiques est la clef de voûte de l'analyse des systèmes. C'est l'étape centrale de l'étude ou de l'élaboration de tout système. De la qualité du modèle dépend toute la suite de l'analyse.

Certes, en recherche opérationnelle comme en théorie de la commande, la construction du modèle a toujours occupé une place importante. Mais ce n'est que dernièrement, avec l'apparition de l'analyse des systèmes qui opère avec des processus reliant des événements de nature physique diverse, qu'est apparue la nécessité pratique d'étudier plus profondément les principes de simulation.

La construction des modèles est toujours une procédure non formelle qui bien évidemment, dépend fortement de l'analyste, de son expérience, de son talent, qui s'appuie toujours sur un certain matériel expérimental, ce qui nous fait dire que le processus de simulation a une base phénoménologique. Le modèle doit reproduire assez fidèlement le phénomène étudié, mais cela n'est pas tout. Il doit encore être facile à manipuler. Donc, la minutie de la description du modèle, sa forme de représentation sont définies par les

objectifs de la recherche et dépendent directement de l'analyste. Des analystes travaillant sur un même système sortiront des représentations différentes.

L'étude et la formalisation du système réel ne constituent pas la seule méthode de construction du modèle mathématique. Il est important d'obtenir des modèles décrivant des phénomènes à partir de modèles simulant des phénomènes plus généraux. Ainsi, le modèle de couche limite de Prandtl peut être déduit à partir du modèle plus général des équations de Navier-Stokes. C'est un modèle asymptotique. Les nouveaux résultats peuvent déboucher sur un modèle plus perfectionné qui rend le précédent modèle asymptotique. Il en fut ainsi de la mécanique newtonienne qui longtemps fut un modèle phénoménologique pur. La mécanique newtonienne est un corollaire de la théorie de la relativité spéciale et elle en résulte par le passage à la limite $v^2/c^2 \to 0$, où $v$ est la vitesse du mouvement propre, $c$, la vitesse de la lumière. Les exemples de cette nature sont nombreux. Ils illustrent un important aspect de l'évolution des sciences naturelles. L'apparition d'un grand nombre de modèles « asymptotiques » plaide pour la maturité de la discipline scientifique, pour les liens logiques étroits entre les divers phénomènes abordés dans le cadre de cette discipline.

La description mathématique et la construction de modèles mathématiques couvrent aujourd'hui des domaines scientifiques extrêmement vastes et de nombreux principes et approches ont été mis au point qui revêtent un caractère assez général dans les conditions actuelles. Voyons ceci de plus près.

Le problème fondamental de l'analyse scientifique est de dégager les mouvements *) réels d'un ensemble de mouvements permis, de formuler les principes régissant leur sélection. Le problème de la simulation mathématique est de décrire ces principes de sélection dans les termes et les variables qui caractérisent le mieux l'objet étudié. Les principes de sélection retrécissent l'ensemble de mouvements permis en rejetant ceux qui ne peuvent être réalisés. Plus le modèle est parfait, plus l'ensemble des mouvements réels est étroit et plus les prévisions sont exactes. Les principes de sélection diffèrent d'un domaine scientifique à l'autre. Depuis le siècle dernier on distingue trois niveaux d'organisation de la matière : la matière inerte, la matière vivante, et l'organisation suprême de la matière — la matière pensante — la société. Cette division est justifiée par des principes différents de sélection des mouvements réels.

A l'échelon inférieur, c'est-à-dire au niveau de la matière inerte, les principes fondamentaux de sélection sont les lois de conserva-

*) Ici et dans la suite le terme « mouvement » est employé dans une acception large : il désigne tout changement, toute interaction d'objets matériels.

tion de la matière, de la quantité de mouvement, de l'énergie, etc. Toute simulation doit commencer par le choix des variables de phase qui serviront à décrire les lois de la conservation. Mais les lois de la conservation ne définissent pas un mouvement unique et ne constituent pas les seuls principes de sélection. Il est nécessaire de faire la part du second principe de thermodynamique, des principes du minimum de dissipation de l'énergie, de stabilité. Il est très important de poser des conditions (contraintes): initiales, aux limites ou autres.

Les lois de conservation sont dans un certain sens « absolues », mais elles sont en nombre insuffisant. Elles n'assurent pas l'unicité des mouvements permis. Les autres principes de sélection rétrécissent davantage l'ensemble des mouvements permis.

Le principe du minimum de dissipation de l'énergie sélectionne parmi les mouvements permis obéissant aux lois de conservation ceux dont la réalisation entraîne un accroissement minimal de l'entropie. Le principe de stabilité conduit l'analyste à porter son attention sur l'étude des seules formes de mouvement dont le temps caractéristique d'existence est suffisamment élevé, etc.

Il existe une certaine « hiérarchie » entre les divers principes de sélection. L'étude de la turbulence montre par exemple que seule existe sa forme stable qui conduit simultanément à une vitesse de croissance minimale de l'entropie.

La construction du modèle et sa mise au point s'appuient sur les principes de sélection et à leur tour en forment de nouveaux, c'est-à-dire des lois dans le cadre desquelles le processus étudié doit toujours rester. Tous les principes de sélection des mouvements, valables pour la matière inerte, restent en vigueur pour la matière vivante *). Donc, même ici la simulation commence par l'écriture des lois de conservation. Mais les variables essentielles sont différentes.

Supposons par exemple qu'il s'agisse d'un macrosystème biologique. Les processus qui s'y déroulent sont des processus d'existence de communautés d'espèces biologiques. La première étude systématique de la dynamique de ces systèmes a été entreprise par le mathématicien italien V. Volterra. La caractérisation essentielle de ces systèmes est la nutrition. Donc, les lois de conservation de la ma-

*) Le fait que les lois valables pour la matière inerte le soient pour la matière vivante a été longtemps controversé. Les plus grandes difficultés ont été posées par le second principe de thermodynamique. Ce problème a été résolu dans les années trente par L. Bertalanffy qui apparement a le premier démontré que les êtres vivants sont des systèmes ouverts, c'est-à-dire qu'ils ne peuvent exister sans échange de matière et d'énergie avec le milieu ambiant (ceci explique la diminution de l'entropie observée chez eux). Ces recherches constituent le principal apport de Bertalanffy à la biologie et à la théorie des systèmes (cf. [21]).

tière et de l'énergie doivent être exprimées en termes trophiques: qui mangera qui et en quelle quantité. L'ouvrage [14] est consacré à la déduction de ces relations et à leur étude.

Mais les principes de sélection des mouvements réels, qui sont inhérents à la matière inerte, ne suffisent pas à eux seuls pour expliquer le contenu des processus qui se déroulent dans la matière vivante. Les mouvements des organismes vivants (mouvements dont la sélection obéit aux lois de la matière inerte) ne sont pas une conséquence des lois de la conservation qui régissent les processus se déroulant dans la matière inerte. La situation se complique ici par le fait que la matière vivante est caractérisée par un comportement rationnel, c'est pourquoi il est pratiquement impossible d'expliquer les choses observées dans le monde vivant sans se servir des notions de rétroaction et d'information.

Dans la suite on utilisera souvent l'« homéostasie ». Il existe plusieurs définitions de ce terme qui diffèrent sensiblement l'une de l'autre. On conviendra d'appeler domaine d'homéostasie d'un organisme (ou domaine de stabilité) le domaine des paramètres exogènes (les paramètres du milieu) dans lequel cet organisme peut survivre.

Tout organisme vivant vise à conserver sa stabilité (son homéostasie). Cela signifie que si les conditions extérieures se modifient il doit se conduire de telle sorte que son état reste dans le domaine des paramètres qui assure sa survie. Tout organisme vivant est doté de récepteurs, qui lui permettent de se situer par rapport à la frontière du domaine d'homéostasie (le vecteur $x$), et de prendre les mesures adéquates (le vecteur $u$). Ainsi, dès qu'il reçoit une information (un signal) sur le milieu ambiant, il définit sa politique en fonction du caractère de cette information. Cela signifie que les actions de l'organisme vivant, c'est-à-dire les mouvements réels, sont choisis d'une manière cohérente — à l'aide de la rétroaction

$$u = f(x) \tag{2.1}$$

l'organisme essaye de s'éloigner de la frontière du domaine d'homéostasie. L'organisme vivant adopte donc une attitude conséquente: il est capable de modifier sa position par rapport à la frontière du domaine d'homéostasie, de faire varier ses caractéristiques internes dans des limites définies et partant la structure du domaine d'homéostasie. Quelquefois même l'organisme peut modifier les caractéristiques du milieu ambiant.

Le désir de conservation de l'homéostasie engendre des mécanismes bien définis de sélection du comportement qui ne sont pas déduits des principes gouvernant les processus de la matière inerte.

Les tentatives d'expliquer les phénomènes observés dans le monde vivant à partir d'un point de vue physique (ou chimique) relèvent d'un problème appelé problème de réductionisme. Ce pro-

blème n'est pas nouveau et les travaux de Bertalanffy ont fortement contribué à en faire progresser la résolution. Ils ont pour le moins montré que les lois de la physique ne pouvaient être tenues à l'écart de l'étude des processus de la matière vivante et rien de plus. Le problème de réductionisme reste entier. Et dans ces questions nous sommes contraints de suivre les principes de V. Vernadski qui en étudiant les problèmes de l'évolution de la vie sur Terre et l'influence de cette évolution sur celle de la Terre en tant que système n'a pu de toute évidence négliger le problème de l'origine de la vie. Et sans doute il a mieux que quiconque appréhendé toute la complexité de ce problème. Prenant conscience de son impuissance à étudier ce problème, Vernadski proposa de l'« excommunier », de l'exclure du cadre de ses recherches. Il jugea utile de constater l'existence de la vie, sans plus. Grâce à cette restriction artificielle de l'objet de recherches, grâce à l'introduction de ce postulat, Vernadski réussit à édifier l'une des plus remarquables théories: la biogéochimie. Nous nous trouvons à peu de choses près dans la même situation. Nous ne pouvons que constater la présence de mécanismes qui sont vitaux à l'existence et au fonctionnement des macrosystèmes biologiques, mais quant à la question de savoir comment ces mécanismes sont liés aux principes de sélection, aux lois établies en physique et en chimie, nous admettrons qu'elle sort du cadre de l'analyse des systèmes (ou plus exactement des questions dont nous nous occupons).

Ainsi, nous ne pouvons décrire le fonctionnement d'un système vivant sans la rétroaction. Signalons encore une fois que les relations (2.1) sont dites rétroactions dans le cas seulement où elles ne peuvent être déduites des lois de physique. Cette circonstance n'est pas toujours correctement comprise et le terme de rétroaction est souvent utilisé en analyse des systèmes pour désigner aussi des relations qui peuvent résulter par exemple des lois de conservation. Ici et dans la suite on comprendra par rétroactions des relations qui revêtent un caractère rationnel ou dirigé. Il est donc absurde de se servir de la rétroaction pour décrire un phénomène se déroulant dans le monde inerte: on ne doit utiliser un terme que lorsqu'on est dans l'impossibilité de s'en passer.

REMARQUE. La notion de rétroaction est apparue en technique. Son usage ici est pertinent, car les systèmes techniques sont une création de l'homme. Ils peuvent être considérés comme une matière inerte créée par l'homme. Donc, les notions d'information, de rétroaction se prêtent bien à la description des systèmes techniques. Ainsi, la structure de la rétroaction réalisée par le pilote automatique n'est pas une conséquence des lois de conservation mais une idée du constructeur.

Les systèmes biologiques se rapportent à la classe des systèmes commandés réflectifs. Ces systèmes sont commandés parce qu'ils contiennent des fonctions qui sont utilisées pour réaliser les objectifs

poursuivis, réflectifs, parce que les fonctions de comportement le sont. Le terme réflectif souligne la simplicité de la dépendance de la commande par rapport à l'information (le réflexe consécutif à l'excitation). Ce terme a été introduit par les biologistes et en premier lieu par l'école de I. Pavlov. Nous l'employerons dans son acception liminale.

La notion d'organisme joue un grand rôle dans la description du fonctionnement des formes biologiques d'organisation de la matière. On appelle organisme tout système possédant des objectifs propres et les moyens (les ressources) pour les atteindre. Seul un organisme est capable de créer la boucle de rétroaction.

Tout individu est un organisme. C'est une évidence. Les groupes d'animaux par exemple présentent certaines caractéristiques des organismes. Dans certaines conditions, une population peut être traitée comme un organisme [24]. Tout macrosystème biologique est un système hiérarchique. L'analyse montre que les objectifs de l'échelon supérieur de la hiérarchie ne sont généralement pas identiques à ceux du niveau inférieur. Par exemple, les intérêts d'un troupeau (son homéostasie) ne sont pas les mêmes que le désir de chaque animal de conserver son homéostasie. Les niveaux hiérarchiques supérieurs ne peuvent plus être considérés comme des organismes. La biogéocénose (système écologique) par exemple n'est probablement pas un organisme même si elle possède une homéostasie. Mais elle n'a apparemment pas la puissance nécessaire pour organiser la boucle supérieure de rétroaction, c'est-à-dire pour utiliser adéquatement ses ressources à la conservation de son homéostasie. Or cela ne veut nullement dire qu'il est possible de décrire la dynamique d'un écosystème sans se servir de la notion de rétroaction, car ce système est composé d'un grand nombre d'organismes.

L'impossibilité de décrire le fonctionnement de tout organisme vivant sans utiliser la notion de rétroaction est un fait notoire depuis assez longtemps, il est même antérieur à l'apparition de cette notion. En tout cas ce fait était connu de A. Bogdanov en 1911. Le principe de rétroaction a été formulé dans les termes actuels en 1931 par le fondateur de la biocybernétique, P. Anokhine.

REMARQUE. Donc, les allégations de N. Wiener en vertu desquelles il aurait introduit le principe de rétroaction dans la théorie des systèmes biologiques nous semblent pour le moins injustifiées.

Ainsi, pour décrire des macrosystèmes biologiques, nous devons nous baser sur les lois de conservation et sur les rétroactions qui sont souvent appelées fonctions de comportement.

Au niveau social de l'organisation de la matière, on a affaire à un phénomène totalement nouveau : le travail. C'est pourquoi nous devons décrire les modèles de ce domaine en termes de travail (en termes d'économie). Voyons à titre d'exemple les relations d'équilibre qui sont classiques en économie.

Désignons par $x$ le vecteur des biens produits. Ses composantes représentent les quantités des divers liens produits: par exemple $x^1$ est la quantité d'acier fondu, $x^2$, la quantité de métaux non ferreux, $x^3$, la quantité de machines-outils, etc. Soit $A = (a_{ij})$ la matrice des dépenses directes, c'est-à-dire que $a_{ij}$ représente la quantité de bien $j$ nécessaire à la production d'une unité de bien $i$. Dans ces conditions on a l'équation d'équilibre évidente:

$$x = Ax + y,$$

ou en coordonnées

$$x^i = \sum_j a_{ij} x^j + y^i, \quad i, j = 1, \ldots, n. \tag{2.2}$$

Le vecteur $y = \{y^1, \ldots, y^n\}$ s'appelle *demande finale* (la production finale peut être investie, consommée, stockée, etc.).

Les relations (2.2) représentent un modèle économique élémentaire appelé modèle de Léontieff (du nom de l'économiste américain V. Léontieff qui le premier dès les années trente commença à utiliser de tels modèles). Ce modèle ne fait intervenir que les lois de conservation (les relations d'équilibre). De même que les modèles de Volterra, les modèles économiques d'équilibre décrivent des flots de matières: des biens matériels, des produits. Il existe actuellement une théorie bien élaborée pour de tels modèles.

Quand on décrit des processus biologiques, on parle d'actions rationnelles. Il est évident que dans les processus sociaux, on a affaire à des actions dirigées. Ici aussi nous parlerons des rétroactions, des processus d'information qui sont infiniment plus compliqués que dans le cas de la matière inerte. Signalons que nous n'avons pas besoin du terme information pour décrire les processus se déroulant dans la matière inerte. Ce terme ne se présente que quand il est question d'actions rationnelles ou dirigées, c'est-à-dire au niveau seulement de la matière vivante.

Les systèmes biologiques mettent en jeu des processus d'information très simples. Les fonctions de comportement réflectif, qui sont décrites par des relations fonctionnelles simples:

$$\text{réaction} = f\ (\text{signal})$$

sont en fait une paramétrisation des processus d'information qui se déroulent dans les systèmes biologiques. La situation est différente pour les processus sociaux. En les simulant nous devons souvent décrire spécialement les procédures de traitement de l'information, et ceci pas seulement du fait que ces processus peuvent être assez longs et complexes à cause de la masse d'information circulant dans le système. L'essentiel est probablement ailleurs. L'homme prend une décision sur la base de l'information reçue et la relation signal-réaction ne revêt plus un caractère de réflexe. Premièrement, c'est

un opérateur complexe et, deuxièmement, elle n'est pas univoque. Nous sommes dans l'obligation de prendre en considération le facteur subjectif. Ces circonstances n'épuisent pas les difficultés posées par la construction des modèles décrivant le fonctionnement des collectivités humaines. Nous avons signalé plus haut le rôle de la notion d'organisme et le fait qu'un organisme ne désigne pas un seul être vivant, mais parfois dans certaines conditions un groupe d'êtres, voire même toute une population. Mais ces macrosystèmes épuisent pratiquement la notion d'organisme au niveau purement biologique. Il en va autrement de la société humaine. Ici, tout groupe de personnes, toute collectivité possèdent des objectifs personnels et les moyens de les réaliser. (Les communautés liées à l'activité industrielle prennent une importance particulière.) Les intérêts des divers groupes peuvent fortement diverger et parfois même devenir antagonistes. Pour décrire un processus social plus ou moins adéquatement nous devons nécessairement apprendre à décrire l'éventail excessivement complexe des intérêts et contradictions sociales. La théorie classique de l'économie politique est un brillant exemple d'unification des principes purement économiques et sociaux, une unification sans laquelle il est probablement impossible de comprendre la quintessence des processus sociaux et de construire un modèle doué de bonnes qualités prévisionnelles.

Il est clair que les intérêts (les objectifs) des divers groupes sont liés à leur homéostasie, c'est pourquoi il y a lieu de parler ici de communautés homéostatiques. Mais les liens qui rattachent les conditions de stabilité et les objectifs qui ont une incidence directe sur le caractère des décisions prises par les organismes aux mesures adoptées pour réaliser ces objectifs sont généralement assez complexes. On est loin d'un pur comportement réflectif. L'homme a le don d'analyser la réalité ambiante, de prévoir l'issue de ses actions, d'avancer des hypothèses sur le comportement des autres personnes, etc. Donc, les rétroactions qui apparaissent dans la société humaine ne peuvent être réalisées par des fonctions élémentaires réflectives. La description mathématique de la non-réflectivité est un problème extrêmement compliqué. Très souvent ou plus exactement en principe nous ne sommes pas en mesure de formaliser les processus de nature sociale et nous devons pour les décrire utiliser une paramétrisation obtenue sous forme de fonctions de comportement à l'aide d'estimations. Les indéterminations posées par la description de ces processus nécessitent la création d'une technique spéciale évoquée dans les chapitres précédents, technique à laquelle est consacrée la majeure partie de cet ouvrage.

Malgré toutes ces difficultés, la description mathématique, c'est-à-dire la simulation mathématique s'est transformée en une discipline scientifique élaborée. Il est évident que le rôle des modèles varie avec le domaine scientifique considéré. Si en physique et en techni-

que les modèles mathématiques constituent l'une des principales méthodes de recherche et d'élaboration des projets, dans les macrosystèmes sociaux et biologiques, ils servent moins à obtenir des caractéristiques chiffrées qu'à établir des estimations délimitant les bornes de nos actions ou les possibilités des processus envisagés et les tendances de leur développement. Grand est le rôle des modèles mathématiques en tant que langage de description permettant de structurer et de réglementer les efforts des chercheurs.

En conclusion, considérons une classification conventionnelle des modèles mathématiques établie actuellement suivant le caractère et la méthode d'utilisation des fonctions et paramètres arbitraires mis en jeu.

a) *Modèles sans commande.* Ces modèles décrivent des processus dynamiques (à l'aide par exemple d'équations différentielles ou aux différences) ne contenant pas de paramètre ou de fonction libres. C'est le cas notamment de la plupart des modèles de prédiction pure dans lesquels l'état initial donné définit la trajectoire du processus. Les modèles de cette nature peuvent être stochastiques, par exemple contenir des variables et des fonctions aléatoires:

$$x = f(x, t, \xi),$$

où $\xi$ est un vecteur aléatoire de loi de probabilité connue. Dans ce cas, on s'intéresse non pas à des trajectoires isolées, mais à leurs propriétés stochastiques, par exemple à la valeur moyenne.

Ces modèles décrivent essentiellement des processus se déroulant dans la matière inerte.

b) *Modèles utilisables pour l'optimisation de certaines activités.* Considérons un processus dynamique dont le modèle est régi par une équation de la forme

$$x = f(x, t, u), \tag{2.3}$$

où le choix de la fonction $u(t, x)$ est du ressort d'un responsable. La fonction vectorielle $u(t, x)$ s'appelle commande. La commande se déduit à partir d'une condition de réalisabilité d'un objectif. Une classe assez répandue de problèmes peuvent être décrits par ces modèles de la manière suivante: transférer en un temps $T$ un système de l'état

$$x(0) = x_0 \tag{2.4}$$

à l'état

$$x(T) = x_T \tag{2.5}$$

de façon à minimiser les « dépenses », c'est-à-dire que

$$\int_0^T F(x, u, t)\, dt \Rightarrow \min. \tag{2.6}$$

Les contraintes (2.4), (2.5) et la fonction objectif (2.6) ne sont pas incluses dans la notion de modèle. Pour un même modèle (2.3), on peut poser des problèmes différents.

Les modèles de cette nature ont fait l'objet du chapitre précédent.

c) *Modèles utilisables pour l'analyse des situations conflictuelles.* Supposons qu'un processus dynamique soit défini par les actions de plusieurs personnes disposant des commandes $u$, $v$, $w$, . . . Alors

$$\dot{x} = f(x, t, u, v, w, \ldots), \tag{2.7}$$

les commandes étant déduites à partir des conditions

$$\int_0^T F_1(x, u, v, w, \ldots, t)\, dt \Rightarrow \min,$$

$$\int_0^T F_2(x, u, v, w, \ldots, t)\, dt \Rightarrow \min,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

traduisant chacune des intérêts bien définis des divers sujets.

L'analyse de tels modèles nécessite la création d'un appareil spécial. Ces modèles décrivent une classe de systèmes dits cybernétiques qui seront examinés dans le prochain paragraphe.

Mais les modèles décrits ne couvrent pas un grand nombre de situations dont la nécessité de l'étude a conduit à la création de l'analyse des systèmes: il s'agit des situations qui ne peuvent être entièrement formalisées et pour l'étude desquelles il est nécessaire d'inclure dans le modèle mathématique un maillon « biologique »: l'homme (l'expert).

## § 3. Systèmes cybernétiques

Au paragraphe précédent nous avons introduit les systèmes cybernétiques comme une classe de systèmes généralisant les systèmes commandés. Contrairement au système commandé auquel on a supposé qu'est associé un responsable (ou un analyste), au système cybernétique est associé tout un groupe de personnes poursuivant des objectifs propres.

REMARQUE. Le terme cybernétique remonte à l'Antiquité et son histoire est longue et complexe. Dans la Grèce Antique existait la notion de *kuberno*, objet de commande composé d'un ensemble de personnes visant des objectifs propres. Ce mot est passé dans de nombreuses langues européennes pour donner « gouvernement » en français, « goubernia » (province) en russe, etc. Tout village, toute ville ou arrondissement avec leur population est un *kuberno*. Un vaisseau pris comme une construction technique n'est pas un *kuberno*, mais il le devient aussitôt qu'il est considéré avec son équipage et ses passagers. L'administrateur d'un *kuberno* est un *kubernet* (« goubernator » en russe et « gouverneur » en français). La cybernétique s'est présentée comme élaborant des recommandations

pour le *kubernet* sur la façon de diriger son *kuberno*. Ces questions ont attiré à des époques différentes de nombreux penseurs universels. Au XIXe siècle d'importants travaux furent publiés en France par Ampère, en Allemagne et en Pologne par Trentovski. Au début des années quarante du siècle passé, B. Trentovski fit à l'université de Fribourg-en-Brisgau un cours spécial sur la philosophie de la cybernétique, cours sur la base duquel il édita à Poznan en 1843 une monographie avec le même intitulé. La signification de ce terme s'est légèrement modifiée au XXe siècle. En introduisant le terme de *systèmes cybernétiques* je ressuscite dans une certaine mesure la tradition classique, car j'envisage un système avec plusieurs sujets poursuivant chacun des objectifs personnels et ayant les moyens de les réaliser.

Quand on parle d'un système gouverné, la principale de ses caractéristiques qu'on a à l'esprit est l'existence de fonctions libres que le responsable associé au système peut utiliser dans ses propres intérêts. Le système peut être dynamique (régi par des équations différentielles ou des équations aux différences) ou statique, mais gouvernable par essence. De même, un système cybernétique se distingue tout d'abord par l'existence de nombreux sujets doués chacun de la possibilité d'influer sur le système tout entier, de modifier le caractère du mouvement de ce dernier dans ses propres intérêts. Ceci est essentiel. Nous étudierons les systèmes cybernétiques le plus souvent par leurs équations différentielles ou en traitant un cas particulier de problèmes statiques. Mais beaucoup des faits établis sont valables au cas où l'évolution des systèmes est régie par des équations aux différences.

Ainsi, on appellera *système cybernétique* un système (2.7)

$$\dot{x} = f(x, t, u, v, w, \ldots, \xi), \tag{3.1}$$

où $u, v, w, \ldots$ sont les commandes des divers sujets, mais contrairement à (2.7) nous considérons le cas plus général où le second membre de (3.1) renferme une fonction vectorielle aléatoire $\xi(t)$.

Pour étudier les systèmes commandés on s'est toujours servi d'une description qui traduisait le niveau des connaissances de l'analyste et du responsable, c'est-à-dire d'une description toujours subjective. Mais comme le modèle ne fournit qu'une description approximative, en l'identifiant au processus copié, on se livre à une conjecture: on admet l'homologie du modèle et du processus étudié.

Au système cybernétique est associée toute une série de sujets avec leurs propres idées sur ce système, des idées qui sont loin d'être identiques. Donc, les systèmes cybernétiques ne peuvent à plus forte raison faire l'objet d'une « description objective » et leur étude ne peut être conduite que du point de vue d'un sujet (ou d'un groupe de sujets) bien déterminé et en fonction de ses objectifs et de son interprétation de la situation.

Signalons que dans les systèmes réels de cette nature, il n'existe

pas de maillon (d'élément) concentrant toute l'information « objective » sur le système.

En étudiant les systèmes commandés, on a attiré l'attention sur le fait que le responsable doit dans des conditions d'indétermination adopter des hypothèses sur la situation, les forces en jeu, etc. Ces hypothèses se compliquent énormément dans le cas de systèmes cybernétiques. Pour calculer par exemple une trajectoire du système, nous devons non seulement nous donner un modèle de la situation, c'est-à-dire formuler des hypothèses sur la nature des fonctions vectorielles $\xi(t)$, mais aussi conjecturer les valeurs des commandes des autres sujets. Quels points de repère avons-nous pour conjecturer le comportement des autres sujets?

Nous partirons du fait que chaque sujet vise un but objectif. Et si un tel but existe, on peut en général le formuler en termes de maximisation d'une fonctionnelle. Nous écrirons le but du sujet $i$ sous la forme

$$J_i \Rightarrow \max.$$

Le fait est qu'en règle générale nous ne connaissons pas exactement ce but. Plus, le sujet $i$ lui-même peut ne pas le connaître (qu'on se rappelle les raisonnements faits sur l'indétermination des objectifs).

Supposons maintenant qu'on ait formulé une hypothèse sur la nature des buts d'un sujet. Ceci est insuffisant. Les actions de ce sujet, c'est-à-dire les valeurs de ses commandes, dépendront encore de son degré d'information. De plus, il est très important de savoir ce que les autres sujets du système cybernétique pensent des objectifs du sujet qui conduit l'analyse, de savoir sur la base de quelle information ils prennent leurs décisions, de savoir aussi ce qu'ils savent sur le degré d'information de ce sujet, etc. Donc, à la différence des systèmes commandés ordinaires, le choix de « notre » commande, par exemple $u_1(t)$, dépendra non seulement de « notre » objectif

$$J_1 \Rightarrow \max$$

et de la situation (c'est-à-dire de $\xi(t)$), mais aussi d'autres facteurs, de leur impact sur le résultat définitif de la décision prise, etc.

A l'heure actuelle on étudie de nombreux types de systèmes, mais il est encore prématuré de mettre au point une théorie passe-partout. Dans ce contexte il est important de dégager une classe de systèmes justiciables d'approches plus ou moins générales et aux propriétés générales descriptibles. Nous aborderons quelques systèmes répondant plus ou moins à ces exigences, il s'agit des systèmes à structure hiérarchique et des systèmes hermeyeriens.

a) *Justifications de l'introduction des structures hiérarchiques.* La hiérarchie suppose une certaine « inégalité », une subordination de certains éléments du système à d'autres. Les notions d'« inégali-

té », et de « subordination » appellent bien sûr des précisions et des commentaires.

Le terme « organisation hiérarchique » (ou « structure hiérarchique ») est utilisé dans des contextes assez variés. Les systèmes hiérarchiques sont largement implantés en technique : les systèmes complexes de liaison, de traitement des données, de gestion du transport, etc., sont toujours organisés d'après un principe hiérarchique qui permet de réaliser simultanément les différentes opérations, de travailler avec des massifs d'information séparés, etc.

La hiérarchisation des systèmes techniques est la conséquence de leur complexité : en effet, le traitement centralisé de l'information soit est impossible, soit implique une dépense de temps (ou de ressources) qui est inadmissible pour des raisons techniques.

Le terme « hiérarchie » est combiné avec celui de « commande » : exemple, « système de commande hiérarchique ». L'objectif essentiel d'une organisation hiérarchique est la répartition des fonctions de traitement de l'information et de prise de décision entre les divers éléments du système. Si le volume de l'information nécessaire à la prise de décision est peu élevé, il n'est point besoin de mettre en place un système de répartition des obligations pour la prise des décisions : celles-ci peuvent être appliquées de manière centralisée.

Toute structure sous-entend des contraintes supplémentaires qui rétrécissent dans le cas général l'ensemble des stratégies tolérables. Désignons par $f(x)$ la fonction objectif du système. Il est alors évident que

$$\max_{x \in G'} f(x) \leqslant \max_{x \in G} f(x), \tag{3.2}$$

pourvu que $G' \subset G$.

Donc, le renoncement à un système de commande entièrement centralisé (le rétrécissement volontaire de l'ensemble $G$ des stratégies) doit être justifié par d'autres circonstances. A cet effet il faut notamment étudier avec plus de détails la structure des processus d'information et la dépendance de la qualité de l'information par rapport à l'organisation du système.

Soient $f(u, \xi)$ la fonction objectif à maximiser, $u = \{u_1, \ldots, u_m\}$, la commande, $\xi$, une variable caractérisant les indéterminations. Supposons que dans des conditions de totale centralisation $u \in G_u$ et $\xi \in G_\xi$. Donc, l'espérance de la fonction objectif sera de la forme

$$f^* = \max_{u \in G_u} \min_{\xi \in G_\xi} f(u, \xi). \tag{3.3}$$

Si maintenant nous introduisons une structure hiérarchique dans le système, cela veut dire que nous répartissons la fonction de commande sur les divers échelons. En d'autres termes, les diverses

décisions seront prises sur la base d'une information partielle. Dans un système de gestion du transport, par exemple, nous avons affaire à une centralisation totale lorsque la décision — pour fixer les idées, l'établissement d'un horaire — est prise en tenant compte de la situation sur toutes les lignes et toutes les gares. Nous pouvons en principe procéder à ces calculs et composer un « horaire optimal », mais cette procédure exigerait un temps si élevé qu'à sa sortie il serait périmé. C'est pourquoi nous sommes dans l'obligation d'établir des horaires pour des tronçons de réseau ferroviaire en n'utilisant que la partie de l'information qui les concerne. Donc, en désagrégeant l'information pour la répartir sur les divers échelons, on peut prendre une décision en négligeant des dépendances plus complexes, par exemple la situation sur la ligne voisine. Nous passons donc à un ensemble plus étroit de stratégies $G' \subset G$. Mais nous constatons en même temps une réduction du niveau d'indétermination. Si le traitement de l'information est décentralisé, certains massifs peuvent être traités plus en détail, c'est-à-dire qu'on peut réduire le niveau des indéterminations et améliorer la qualité de l'information. Donc, $\xi \in G'_\xi \subset G_\xi$ et l'on obtient une autre espérance

$$\dot{f} = \max_{u \in G'_u} \min_{\xi \in G'_\xi} f(u, \xi). \tag{3.4}$$

L'adéquation de l'introduction d'une structure donnée dans le système de gestion est finalement mise en évidence par la comparaison de quantités de type (3.3) et (3.4).

Donc, l'estimation de l'organisation hiérarchique se ramène à la comparaison de deux tendances contradictoires. Le passage à une structure hiérarchique rétrécit l'ensemble des stratégies et baisse simultanément le niveau d'indétermination, c'est-à-dire rend possible l'acquisition d'une solution meilleure.

Le choix d'une structure hiérarchique se heurte à une difficulté. Le nombre de formes architecturales possibles est fini. Quand on projette un système, on doit en même temps choisir une stratégie (définir les valeurs des commandes) et la meilleure « architecture ». Il existe ainsi plusieurs versions de structures hiérarchiques possibles. Nous disons qu'il existe un ensemble discret de structures $S$ ou si l'on veut $S$ est l'ensemble des projets. A chaque projet $s \in S$ sont associés un seul ensemble de stratégies $u^s \in G^s_u$ et un seul ensemble d'indéterminations $G^s_\xi$. A ces ensembles correspondent les espérances

$$f^s = \max_{u \in G^s_u} \min_{\xi \in G^s_\xi} f(u, \xi)$$

et le problème de l'élaboration d'une structure hiérarchique peut être ramené en fin de compte à la recherche d'un élément $s \in S$,

solution du problème

$$f^s \Rightarrow \max_{s \in S} {}^{*}). \tag{3.5}$$

Nous avons employé l'expression « en fin de compte » pour souligner qu'avant de résoudre le problème (3.5), il faut encore construire l'ensemble $S$. Si la résolution du problème (3.5) est formelle dans une certaine mesure, la construction de $S$, ensemble des « schémas de construction » possibles, relève de l'ingéniosité et n'est pas formelle.

Ainsi pour élaborer un système hiérarchique de commande dans les systèmes techniques, on répartit les fonctions de prise de décision entre les divers échelons du système. En outre, avec la description des « objectifs » on doit formuler (projeter) un certain algorithme, c'est-à-dire des règles de traitement de l'information et des règles de prise de décision sur la base de l'information recueillie. On obtient finalement un système réflectif.

REMARQUE. Signalons que les maillons de ce système sont plutôt des personnes, par exemple des dispatchers. Néanmoins ces personnes doivent se conformer à des règles de comportement bien définies dans les systèmes techniques et tout écart de conduite peut être considéré (à la rigueur) comme un « bruit » supplémentaire accroissant le niveau d'indétermination. Dans les systèmes techniques on peut généralement ne pas tenir compte de la non-réflectivité des maillons « biologiques ». Donc, les « tâches » qui sont réparties entre les divers maillons d'un système peuvent très bien ne pas être accomplies soit que les maillons sont en dérangement (la négligence des opérateurs ou des dispatchers n'est pas exclue), soit à cause de l'apparition de perturbations extérieures imprévisibles.

La situation est totalement différente lorsqu'on élabore des structures hiérarchiques de gestion de systèmes industriels, économiques ou sociaux dans lesquels les principaux objets de gestion sont des personnes et non des machines.

La nécessité de doter les systèmes sociaux d'une structure hiérarchique tient aux mêmes raisons que dans les systèmes techniques: l'impossibilité d'un traitement centralisé de l'information. Cette impossibilité de traiter l'information dans les délais fixés et convenablement (c'est-à-dire de justifier la décision) fait que les décisions ne sont pas bien pesées, ce qui équivaut à un niveau d'indétermination élevé. Force est donc de désagréger le traitement de l'information et de déléguer le droit de décision aux échelons inférieurs.

Ainsi, le directeur d'un trust de sovkhozes d'une grande région ne peut connaître la situation des divers sovkhozes avec les mêmes détails que les directeurs de ces derniers. Donc, les décisions adoptées

---

*) Le schéma d'estimation de la qualité de la structure que nous venons de développer se rapporte aux systèmes techniques (réflectifs). L'extension de ces raisonnements à des systèmes « intelligents » implique, comme nous le verrons plus bas, des hypothèses supplémentaires et une simulation sur machine.

10—01523

par ces directeurs seront plus adéquates (du point de vue des objectifs poursuivis) que celle prise en haut lieu. D'où la nécessité d'une certaine décentralisation dans la prise des décisions, c'est-à-dire d'une décentralisation de la gestion. Mais la décentralisation crée des conditions qui contribuent en principe à une diminution de l'efficacité du système. Tout d'abord, de même que dans les systèmes techniques, on constate une restriction de l'ensemble des stratégies, due au fait que les échelons inférieurs ne travaillent qu'avec une partie de l'information. Par ailleurs, de nouveaux facteurs viennent compliquer l'analyse dans les systèmes sociaux.

En effet, dès qu'une partie d'un organisme se voit impartir le droit de décision, elle en profite pour réaliser les buts qui lui sont objectivement inhérents, c'est-à-dire qu'elle se transforme en un organisme indépendant, d'où d'inévitables contradictions entre cette partie et le tout.

Si dans les systèmes techniques nous dotons un élément du système d'une certaine ressource, nous définissons en même temps son mode d'usage. Ce mode doit, dans des conditions concrètes données qui seront communiquées à une certaine date au dispatcher (ou à l'ordinateur), assurer la plus grande efficacité de l'ensemble du système. Notre tâche en tant que concepteurs du système technique de gestion consiste à déterminer le mode d'usage des ressources (par exemple composer le graphique de fonctionnement d'un barrage en fonction des particularités des crues, particularités qui seront connues au début de ces crues).

La situation est différente dans les systèmes sociaux. Certes ici aussi nous pouvons proposer (projeter) des recettes pour utiliser les ressources disponibles. Mais il ne faut pas oublier que les échelons inférieurs ont acquis tous les attributs d'un organisme et qu'ils utiliseront les ressources reçues pour réaliser leurs objectifs personnels. Donc, la gestion de tels systèmes doit différer de celle des systèmes techniques et tenir compte des traits objectifs inhérents aux systèmes sociaux.

Le problème de l'apparition de buts objectifs au sein de groupes et systèmes sociaux est très délicat. D'une manière ou d'une autre il est lié à l'homéostasie. Mais ces liens sont très embrouillés, ils ont été tissés par plusieurs générations et leur appréhension est loin d'être évidente. Ce problème implique une profonde analyse sociologique et philosophique. Mais dans les situations concrètes auxquelles est confronté l'analyste d'un système bien défini, il est souvent possible d'éviter l'analyse de ces problèmes complexes et de formuler plus ou moins exactement les principaux objectifs. Cela étant, il faut toujours avoir présent à l'esprit les contradictions et aussi le fait que tout objectif est le résultat d'un compromis. L'élaboration même d'une structure hiérarchique dans les systèmes sociaux est toujours une sorte de compromis.

J'ai amplement discuté du rôle de l'information dans la création des systèmes hiérarchiques. Il est possible de citer un autre argument en leur faveur. Les sujets se regroupent en structure hiérarchique, car ils sont « gagnants »: en modérant leurs désirs, ils confortent leur stabilité.

Par exemple, les intérêts (objectifs) d'une entreprise sont toujours le résultat de la résolution de contradictions entre l'échelon supérieur, pour fixer les idées l'association dont fait partie cette entreprise, et l'entreprise elle-même.

A la lumière de tels raisonnements, on peut dans de nombreux cas concrets énumérer avec une précision plus ou moins satisfaisante les objectifs de l'organisme étudié.

REMARQUE. La description des systèmes cybernétiques (notamment la formation des hypothèses sur le comportement des sujets du système) donne inévitablement lieu à des indéterminations d'un niveau élevé. Ce manque de netteté dans la position des problèmes mathématiques doit être pris en considération au moment du choix de la méthode d'analyse.

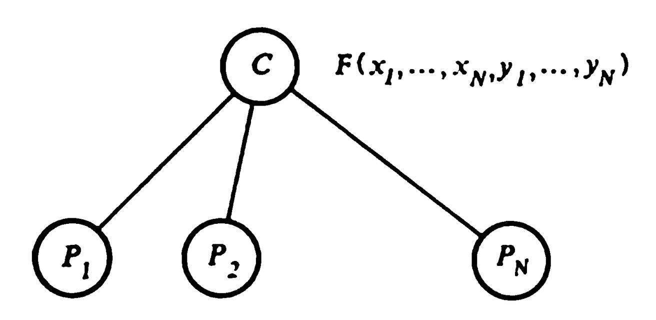



Fig. 3.1

b) *Schémas possibles d'organisations hiérarchiques.* Le schéma le plus simple d'organisation hiérarchique d'un système de gestion est la structure arborescente à deux niveaux montrée sur la fig. 3.1. Ce terme désigne un système mettant en jeu un sujet privilégié, ou responsable, qui a le pouvoir de diriger les autres sujets. Précisons ce terme. Un sujet nommé $C$ (Centre) poursuit des objectifs ainsi notés:

$$F(x_1, \ldots, x_N, y_1, \ldots, y_N) \Rightarrow \max. \tag{3.6}$$

$x_i$ sont les actions du Centre sur des maillons $P_1, \ldots, P_N$ que nous conviendrons d'appeler Producteurs. Les objectifs personnels de ces derniers sont:

$$f_i(x_i, y_i) \Rightarrow \max_{i=1, \ldots, N} \tag{3.7}$$

Donc, les intérêts des Producteurs sont définis par les quantités $y_i$ qui se trouvent à leur disposition et par les quantités $x_i$, qui dépen-

dent du Centre. Dans ce schéma élémentaire, on admet que la valeur de la fonction objectif du Producteur $P_i$ ne dépend pas des actions des autres Producteurs *).

L'inégalité des sujets se manifeste dans le fait que c'est précisément le Centre qui définit les règles de formation des actions $x_i$ qui dépendent d'une manière ou d'une autre de celles des Producteurs (de leurs façons de choisir les $y_i$) et les Producteurs connaissent ces règles au moment où ils décident de choisir les quantités $y_i$. Donc, dans les systèmes hiérarchiques décrits, le Centre a la possibilité (qui s'appelle parfois règle du premier coup ou règle du trait) de canaliser les efforts des échelons inférieurs. Signalons que dans le schéma décrit, le Centre communique une certaine information aux échelons inférieurs, ce qui peut lui être profitable.

Dans les systèmes hiérarchiques, par commande optimale on

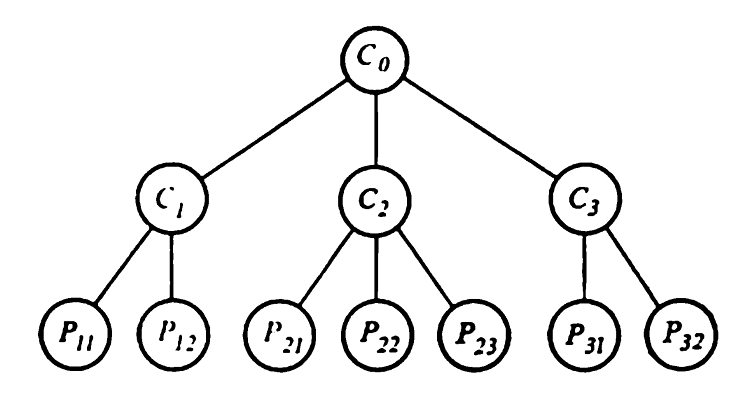



Fig. 3.2

entend une commande $x_i$ qui maximise la fonction objectif (3.6) pour des valeurs données des fonctions $f_i$ caractérisant les intérêts des divers maillons.

La généralisation naturelle de la hiérarchie à deux échelons est la hiérarchie à plusieurs échelons. La figure 3.2 représente un schéma d'une hiérarchie à trois échelons. La fonction objectif du Centre $C_0$ sera de la forme

$$F = F(x_1, \ldots, x_k, y_1, \ldots, y_k),$$

où $x_1, \ldots, x_k$ sont les commandes du centre $C_0$ (les actions de $C_0$ sur les éléments $C_1, \ldots, C_k$). Les fonctions objectifs de ces derniers seront de la forme

$$f_i = f_i(y_{i1}, \ldots, y_{iN_i}, z_{i1}, \ldots, z_{iN_i}) \qquad i = 1, \ldots, k,$$

où $y_{ij}$ est la commande de $C_i$: son action sur le Producteur $P_{ij}$, c'est-à-dire le Producteur d'indice $j$ subordonné à $C_i$.

*) La hiérarchie arborescente est la plus étudiée de toutes les structures hiérarchiques. Cf. par exemple [4, 6, 41, 42].

Enfin les fonctions objectifs des Producteurs $P_{ij}$ seront

$$\varphi_{ij} = \varphi_{ij}(y_{ij}, z_{ij}), \qquad i = 1, \ldots, k; \qquad j = 1, \ldots, N_i,$$

où $z_{ij}$ sont les actions des Producteurs. Dans cette structure hiérarchique, le trait revient au Centre $C_0$ : c'est lui qui communique aux Centres $C_i$ les règles d'affectation des commandes en fonction de leurs choix (de leurs actions). Le coup suivant revient au Centre $C_i$ ($i = 1, 2, \ldots, k$) qui communique les règles de choix des actions $y_{ij}$.

Dans ces schémas, on admet que l'information est échangée verticalement (c'est-à-dire suivant la voie hiérarchique) : l'échange horizontal (c'est-à-dire entre les éléments d'un même niveau) est exclu. On pourrait aussi envisager des généralisations des structures décrites, donnant lieu à un échange horizontal d'information.

L'étude de la vie courante nous confronte aux formes les plus insolites de liaison hiérarchique, des formes qui reflètent la complexité et l'interdépendance des divers processus sociaux, de la production et de la diversité des intérêts des organismes sociaux et technologiques. Les structures hiérarchiques envisagées reflètent d'importantes particularités de la gestion par secteurs industriels : le secteur industriel est sous l'autorité de l'État et à son tour il se décompose en départements, trusts, entreprises. La hiérarchie en éventail se retrouve aussi dans le fonctionnement d'une armée (d'où elle est probablement passée dans la gestion économique). Certes, dans la réalité elle est bien plus complexe. L'échange d'information entre les échelons inférieurs et la transmission de l'information d'un échelon à un autre sans passer par la voie hiérarchique sont inévitables. Il n'empêche toutefois que les schémas décrits englobent les particularités essentielles d'une vaste classe de formes hiérarchiques d'interaction des sujets dans des systèmes cybernétiques réels. C'est pourquoi leur étude s'est transformée au cours des dix dernières années en un chapitre important des mathématiques appliquées et de l'analyse des systèmes.

Mais les schémas en éventail n'épuisent pas bien sûr toutes les formes de structures hiérarchiques. La structure rhomboïdale a pris une grande importance durant ces dernières années.

Toute entreprise est dans des rapports assez complexes avec les autres membres de l'organisation hiérarchique. Tout d'abord elle doit se soumettre au secteur industriel dont elle fait partie. Ensuite elle est un élément de l'infrastructure de la ville ou de la région où elle est implantée. En d'autres termes elle figure dans un système hiérarchique régional. Cette situation est schématiquement représentée sur la figure 3.3. Ces structures sont dites rhomboïdales. Les flèches indiquent le sens de subordination.

Le niveau supérieur a la possibilité d'exercer son autorité sur le second niveau, c'est-à-dire sur l'administration régionale et sur celle du secteur industriel. Le secteur industriel peut à son tour

influencer le comportement du Producteur mais n'a pratiquement aucune prise sur les décisions adoptées au niveau régional. Il en va de même de la direction régionale. Elle a en effet la possibilité de superviser les décisions prises au niveau des entreprises implantées dans la région et ne peut pratiquement pas intervenir dans les décisions adoptées au niveau sectoriel. S'agissant du Producteur, c'est-à-dire de l'entreprise, sa direction peut toujours prendre une décision dans une situation de conflit: d'un côté ses actes sont limités par

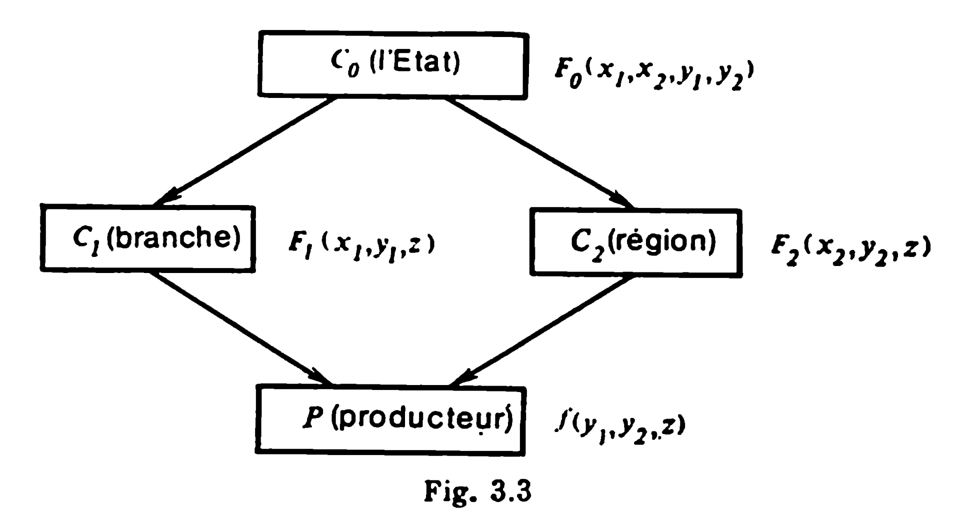



Fig. 3.3

les directions régionale et sectorielle, et de l'autre elle poursuit des objectifs personnels qui sont la conséquence de rapports de production, de normes juridiques et de nombreux autres facteurs déterminant la vie et le développement de la société.

La figure 3.3 ne représente que le schéma le plus élémentaire de structure rhomboïdale. Ainsi, par exemple, l'Etat peut décréter des conditions de fonctionnement d'une entreprise, c'est-à-dire intervenir directement dans les décisions prises au niveau inférieur, certaines structures rhomboïdales peuvent coopérer, etc. Cependant le schéma de la figure 3.3 reflète de nombreux traits importants d'une structure hiérarchique réelle que l'on rencontre souvent dans la pratique. La théorie de la hiérarchie rhomboïdale fait aujourd'hui l'objet de nombreux travaux (cf. par exemple [32]). L'intérêt pour ces recherches est motivé par la position réaliste du problème. La gestion de l'Etat relève nécessairement d'une structure cellulaire, chaque cellule étant une structure rhomboïdale. Un problème d'une grande portée économique est le problème de l'organisation de la gestion suivant le principe régional ou le principe sectoriel. Ces deux schémas ne sont pas contradictoires. Il faudrait combiner de façon rationnelle les deux types de gestion et répartir raisonnablement (voire même optimalement sous certaines conditions) les droits et obligations entre les divers niveaux de la hiérarchie.

Nous avons décrit quelques schémas possibles de systèmes hiérarchiques. Tous ces systèmes sont caractérisés par l'objectivité des intérêts des divers niveaux, par la contradiction d'une partie et du tout. Ceci est un point cardinal de la théorie des systèmes hiérarchiques et est à notre sens primordial pour la compréhension des positions liminales permettant d'étudier les mécanismes de développement des systèmes économiques, écologico-économiques, sociaux et autres.

c) *Systèmes de Hermeyer*. Jusqu'ici nous avons étudié des systèmes cybernétiques à structure hiérarchique, c'est-à-dire dont les protagonistes ne jouissent pas des mêmes droits. Il existe cependant une vaste classe de systèmes cybernétiques dans lesquels tous les sujets sont égaux et d'où est exempte toute hiérarchie. Tel est le cas par exemple de tous les systèmes cybernétiques décrivant des conflits internationaux, l'évolution des conditions écologiques, etc. Le fonctionnement de tels systèmes implique des décisions collectives, donc ils ne peuvent être décrits par des schémas hiérarchiques.

Dans le premier chapitre, nous avons entamé l'étude de systèmes nécessitant des décisions collectives et avons indiqué les difficultés soulevées. Le problème majeur de la théorie de ces systèmes est de trouver des conditions raisonnables de compromis. On a indiqué que l'un des plus importants principes était celui de Pareto : le compromis doit être efficace. Mais l'appartenance à l'ensemble de Pareto ne suffit pas encore à la formation du compromis. Nous avons encore parlé du principe de stabilité de Nash. Si le compromis satisfait le principe de Nash, alors les sujets qui l'ont adopté n'ont pas intérêt à manquer à leurs obligations. Le sujet qui viole la condition de compromis en fait le premier les frais.

Dans l'immense majorité des situations conflictuelles décrites, les compromis stables ne sont pas parétiens et les compromis appartenant à l'ensemble de Pareto ne sont pas stables. Cependant il existe des cas exceptionnels où l'ensemble des compromis efficaces renferme des choix stables. Nous verrons plus loin que ces systèmes sont d'une très grande importance pratique.

Yu. Hermeyer et I. Vatel [35] ont étudié une nouvelle et importante classe de systèmes doués précisément de la propriété indiquée. Ils ont considéré une situation qui en théorie de la recherche opérationnelle a été appelée « voyageurs dans une barque ». Imaginons-nous un système de $N$ partenaires (sujets) égaux poursuivant chacun des objectifs personnels. Mais en plus de leurs propres objectifs, ils ont un but commun : accoster le rivage. Cette circonstance implique de chaque sujet des sacrifices au nom de l'objectif commun.

On considère ainsi un système de $N$ sujets dont les fonctions objectifs seront $f_i(x_i)$, $i = 1, \ldots, N$, où la ressource $x_i$ se trouve à la disposition exclusive du sujet d'indice $i$. Mais en plus de ces objectifs, il existe un objectif commun décrit par la fonction

$F(y_1, \ldots, y_N)$. Les valeurs de cette fonction dépendront de l'activité de tous les partenaires: des ressources $y_i$ consacrées par le sujet $i$ à la réalisation de l'objectif commun, autrement dit sa contribution à l'objectif commun.

Donc, les objectifs de chaque sujet se décrivent par le critère vectoriel

$$f_i(x_i) \Rightarrow \max, \qquad F(y_1, \ldots, y_N) \Rightarrow \max, \tag{3.8}$$

avec $x_i + y_i = a_i$, $i = 1, \ldots, N$, où $a_i$ sont les ressources globales du sujet $i$. Donc, chaque sujet doit dissocier ses ressources pour en consacrer une partie $x_i$ à ses objectifs personnels, « égoïstes », et l'autre partie $y_i = a_i - x_i$, aux objectifs communs.

REMARQUE. Signalons que cette situation est assez typique. De nombreux problèmes d'une grande importance pratique s'y ramènent, par exemple: la réalisation d'obligations communes, les directives administratives, l'affectation des ressources à la protection ou la restauration de l'environnement. L'analyse des variantes possibles de l'attitude de pays souverains pour réaliser un but commun qu'il soit militaire, politique ou économique, fait aussi partie de ces problèmes.

On a vu au premier chapitre que l'analyse d'une telle situation conflictuelle devait commencer par la réduction du critère vectoriel (3.8) à un critère scalaire. Posons

$$J_i = \Psi(f_i(x_i), F(y_1, \ldots, y_N)), \tag{3.9}$$

où $\Psi$ est un opérateur de réduction des critères $f_i$ et $F$, par exemple

$$J_i = \min\{\lambda_i f_i(x_i);\ F(y_1, \ldots, y_N)\} \tag{3.9'}$$

ou

$$J_i = f_i(x_i) + \mu F(y_1, \ldots, y_N), \tag{3.9''}$$

où $\lambda_i$ et $\mu$ sont des coefficients de pondération caractérisant l'intérêt manifesté par les sujets à la réalisation de l'objectif commun. Les systèmes cybernétiques avec des critères (3.9) seront appelés *systèmes hermeyeriens*.

La notion d'homéostasie peut être formulée pour les systèmes cybernétiques hermeyeriens. On appellera frontière du domaine d'homéostasie dans l'espace engendré par les variables $y_1, \ldots, y_N$ la surface $F(y_1, \ldots, y_N) = F_0$ délimitant dans cet espace le domaine d'existence (d'homéostasie, de stabilité) de l'ensemble de tous les sujets. On conviendra qu'au domaine d'homéostasie sont associées les valeurs de la fonctionnelle $F$ supérieures à $F_0$. Donc, la maximisation de $F$ traduit formellement le désir du sujet de se trouver dans l'état le plus stable possible.

Les systèmes hermeyeriens statiques sont justiciables du

THÉORÈME DE HERMEYER-VATEL. *Supposons que $f_i$ et $F$ sont des fonctions monotones strictement croissantes. Il existe alors des*

*solutions stables parmi lesquelles l'une au moins est efficace.*

Les auteurs de ce théorème n'ont pas seulement établi sa véracité, ils ont de plus proposé une méthode efficace de détermination de la situation d'équilibre. Cette méthode se ramène au système de procédures suivant. Ordonnons les sujets comme suit:

$$\lambda_1 f_1 (a_1) \geqslant \lambda_2 f_2 (a_2) \geqslant \ldots \geqslant \lambda_N f_N (a_N).$$

Il existe alors un $p \leqslant N$ tel que tous les $y_i$, $i \geqslant p$, doivent être pris égaux à 0 (c'est-à-dire que les sujets dont les $\lambda_i f_i (a_i)$ sont petits peuvent consacrer toutes leurs ressources à la réalisation de leurs objectifs personnels (égoïstes)). Les autres quantités $y_i$ se déterminent à partir du système d'équations

$$\lambda_i f_i (y_i) = F (y_1, \ldots, y_p, 0, \ldots, 0), \qquad i = 1, 2, \ldots, p.$$

En définitive on trouve la partie des ressources concédées par les sujets à la réalisation de l'objectif commun. Signalons qu'une partie des partenaires ne participent pas du tout aux entreprises collectives, soit qu'ils ne disposent pas d'assez de ressources, soit que leur technologie est d'un faible niveau (les $f_i (a_i)$ sont petits), soit qu'ils sont peu intéressés par les résultats de la décision commune ($\lambda_i$ est petit) (cf. plus en détails [52]).

Dans [35] on ne considère que des systèmes statiques, c'est-à-dire des systèmes ne dépendant pas du temps. La question de savoir dans quelle mesure les résultats obtenus se généralisent aux systèmes dynamiques, reste ouverte.

REMARQUES. 1. Les systèmes hermeyeriens ont acquis de l'importance ces derniers temps en raison des problèmes posés par l'écologie globale. C'est que les systèmes écologico-économiques de caractère global sont nécessairement hermeyeriens. En effet, la rupture de l'équilibre d'un écosystème ou une modification irréversible du climat peuvent sortir l'humanité de son domaine d'homéostasie. Donc, quel que soit l'ensemble de sujets du système cybernétique considéré, parmi les critères qui les guident il existera toujours un critère de stabilité commun à tous ces sujets. Mais si ce système est hermeyerien, alors on a intérêt à chercher les solutions collectives stables appartenant à l'ensemble de Pareto. Donc, ces raisonnements plaident pour l'édification d'une théorie permettant de définir une stratégie industrielle préservant l'environnement. A mesure qu'elle se développera, cette théorie pourra être d'une grande utilité pour la résolution d'innombrables problèmes de gestion des ressources et de l'activité industrielle.

2. Il apparaît très intéressant d'étudier les systèmes hermeyeriens doués d'une certaine structure hiérarchique et dans lesquels les sujets possèdent de plus des critères vectoriels de forme plus générale que (3.8).

d) *Systèmes coopératifs.* La coopération, c'est-à-dire la conjugaison des efforts de plusieurs sujets pour la réalisation d'un objectif commun, est l'une des plus vieilles formes d'organisation de l'activité humaine. Certains systèmes coopératifs peuvent formellement être considérés comme un cas particulier de systèmes hermeyeriens dans la mesure où les sujets d'un système coopératif poursuivent un objectif commun. Cependant certains traits spécifiques de ces systè-

mes nous contraignent à les classer à part. En effet, dans les systèmes hermeyeriens la fonction $F(x_1, \ldots, x_N)$ est supposée donnée et l'analyse de ces systèmes se ramène à la détermination des ressources $x_i$ consacrées à la réalisation de l'objectif commun des sujets coopérants.

Le problème de l'affectation des ressources à la réalisation de l'objectif commun demeure bien sûr dans les systèmes coopératifs. Mais la structure de la fonction $F$ ne peut être considérée comme donnée: elle dépend des clauses de la coopération et aussi bien des quantités $x_i$ que de leur mode d'emploi. La construction de la fonction $F$ est le problème majeur de la synthèse du mécanisme de coopération. On peut la formaliser en introduisant un ensemble $G_p$ de formes possibles d'organisation des systèmes coopératifs. Désignons par

$$F_p(x_1, x_2, \ldots, x_N)$$

la valeur de la fonction objectif « commune » qui aux ressources $x_1, \ldots, x_N$ associe le revenu du système $F_p$ sous réserve que les paramètres de la coopération $p \in G_p$. Dans ces termes, l'objectif « commun » s'écrit:

$$F_p(x_1, x_2, \ldots, x_N) \Rightarrow \max_{x_i}, \quad p \in G_p.$$

Autrement dit, dans les systèmes coopératifs on peut définir toute une classe de fonctions $F_p$, et l'ensemble des stratégies de prise d'une décision collective contient aussi le choix de la fonction $F_p$. Etudions les traits spécifiques des systèmes coopératifs sur un exemple simple.

Supposons que $N$ fermes utilisent l'eau d'un même système d'irrigation pour la production d'un même bien, par exemple du coton. Désignons par $x_i$ la quantité d'eau consommée par la ferme d'indice $i$. La quantité de coton produite alors par cette ferme sera une fonction de $x_i$ et du mode d'usage $c_i$; soit $R(c_i, x_i)$. Si l'on désigne par $q$ le coût d'une unité de bien produit, le revenu brut de la ferme sera $R(c_i, x_i)\,q$. Si chaque ferme fonctionne séparément, alors il n'y a pas d'objectif commun. Si l'on se sert des notations du numéro précédent, on obtient

$$F = F_0(x_1, x_2, \ldots, x_N) = 0 \ \forall x_i$$

et

$$\varphi_i = R_i(c_i, x_i)\,q.$$

Supposons par ailleurs que l'eau est répartie suivant un procédé traditionnel et que chaque ferme n'a le droit d'utiliser qu'une certaine quantité l'eau $x_i^*$. Il ne reste alors à la direction de la ferme

qu'à choisir la méthode la plus rentable

$$\varphi_i^*(x_i, q) = \max_{c_i} R_i(c_i, x_i)\, q.$$

Le revenu total de toutes les fermes est alors

$$F^* = \sum_i \varphi_i^*(x_i, q).$$

En cas de coopération les terres peuvent être utilisées de façon plus rentable. Considérons la somme

$$F = \sum_i \varphi_i^*(x_i, q) \qquad (3.10)$$

où les quantités $x_i$ ne sont pas fixes mais reliées par la relation:

$$x_i \leqslant X, \qquad (3.11)$$

où $X$ est la quantité totale d'eau. En choisissant $x_i$ de manière à maximiser la fonction (3.10) sous la condition (3.11), on s'assure un plus grand revenu global:

$$\max_{\sum x_i \leqslant X} \sum_i \varphi_i^*(x_i, q) = \sum_i \varphi_i^*(\hat{x}_i, q) = \hat{F} \geqslant F^*.$$

La coopération se traduit par un revenu supplémentaire:

$$\Delta = \hat{F} - F^*.$$

Si on divise ce revenu entre les fermes

$$\Delta = \sum \Delta_i$$

de telle sorte que pour chaque $i$

$$\varphi_i^*(\hat{x}_i, q) + \Delta_i \geqslant \varphi_i^*(x_i^*, q),$$

alors chaque ferme gagnera à participer à la coopération. Donc, dans ce cas, on obtient un système dans lequel tous les sujets sont unis par le même objectif: la maximisation du revenu global:

$$F(x_1, x_2, \ldots, x_N) = \sum_i \varphi_i^*(x_i, q) \Rightarrow \max.$$

Il existe une infinité de moyens pour diviser le revenu supplémentaire et tous ces moyens fournissent la même valeur de la fonction objectif $F$. Mais les revenus ne seront pas bien sûr les mêmes pour toutes les fermes.

## § 4. Exemples de systèmes hiérarchiques

Au premier chapitre, nous avons réservé une place importante à l'étude des particularités des couflits entre sujets inégaux. En fait,

nous avions déjà exposé les schémas de raisonnements et la structure des hypothèses généralement utilisés pour l'analyse et l'élaboration des projets de systèmes hiérarchiques de gestion. Dans ce paragraphe, nous nous pencherons sur quelques exemples illustrant les particularités des procédures de prise de décision dans les conflits entre personnes inégales. Au premier chapitre, nous n'avons envisagé que des systèmes statiques. La dynamique, l'évolution dans le temps introduisent de nombreuses particularités dont certaines seront examinées dans ce paragraphe.

a) *Répartition des ressources dans un système hiérarchique économique.* Considérons un groupement de $N$ entreprises industrielles (un trust ou un syndicat) produisant un même bien. Admettons que ce groupement est organisé suivant le principe de la hiérarchie en éventail. A la tête du groupement on trouve un conseil d'administration que l'on appellera Centre. Désignons par $P_i$ le volume de la production de l'entreprise (du Producteur) d'indice $i$. Pour alléger les raisonnements, nous supposerons que les quantités $P_i$ sont scalaires. Le résultat du fonctionnement du Centre est défini par ceux des Producteurs: le Centre ne produit aucun bien. L'activité du Centre peut être estimée de plusieurs manières. Nous glisserons sur leur examen. Ce qui compte pour nous, c'est que la fonction objectif du Centre est définie de façon unique par la production des Producteurs:

$$J = J\,(P_1, \ldots, P_N). \qquad (4.1)$$

Supposons encore que le Centre n'a pas le pouvoir de fixer les volumes de production $P_i$: il ne peut modifier les quantités $P_i$ qu'indirectement, en tenant compte des intérêts et des objectifs des Producteurs. Nous admettrons que le volume $P_i$ de bien produit par le Producteur $i$ est défini de façon unique par les fonds $x_i$ et la main-d'œuvre $L_i$:

$$P_i = f_i\,(x_i, L_i). \qquad (4.2)$$

La fonction $f_i$ s'appelle *fonction de production.*

REMARQUE. La relation (4.2) constitue une hypothèse. La fonction de production de l'entreprise décrit les possibilités extrêmes de la production. A strictement parler, l'égalité (4.2) aurait dû être remplacée par l'inégalité

$$P_i \leqslant f_i\,(x_i, L_i), \qquad (4.3)$$

et le volume de bien produit, traité comme un facteur de gestion soumis à la contrainte (4.3). Pour simplifier les calculs, nous nous bornerons à l'analyse du cas où le Producteur utilise ses possibilités extrêmes.

Il existe plusieurs façons d'approcher la fonction de production. La *fonction de Kobb-Douglas*

$$P_i = \alpha_i x_i^{k_i} L_i^{1-k_i}, \quad k_i \in [0, 1], \quad i = 1, \ldots, N, \qquad (4.4)$$

où $\alpha_i$ et $k_i$ sont des coefficients caractérisant l'entreprise, est largement utilisée dans les recherches en mathématiques de l'économie.

On admettra que le revenu $J_i$ du Producteur $i$ est égal au coût des biens produits moins les frais généraux. Pour simplifier on conviendra que ces frais vont uniquement à la rémunération de la main-d'œuvre. Si l'on désigne par $\omega_i$ le taux de salaire moyen (qui est fixe), alors le revenu $J_i$ sera

$$J_i = c_i P_i - \omega_i L_i, \tag{4.5}$$

où $c_i$ est le prix unitaire du bien $P_i$. Si les fonds sont fixes, le volume de la production est défini de façon unique par la quantité de travail $L_i$. La quantité $L_i$ est un paramètre de gestion qui se trouve à l'entière disposition du Producteur.

Mais pour pouvoir commander les activités du Producteur, le Centre doit être en mesure d'agir sur ses objectifs, objectifs qu'il doit évidemment connaître (ou être supposé connaître). On admettra que chaque Producteur essaye de maximiser son propre revenu:

$$J_i = c_i P_i - \omega_i L_i \Rightarrow \max. \tag{4.6}$$

REMARQUE. Si le Centre ne connaît pas les objectifs du Producteur, il doit alors formuler une hypothèse de type (4.6).

Le Centre peut agir sur le Producteur par le biais des ressources dont il dispose et qu'il doit consacrer à la création des fonds du Producteur. Ainsi, on conviendra que le Centre sait que le Producteur une fois en possession des ressources $u_i$ produira pendant la durée du plan la quantité de biens suivante:

$$P_i = \alpha_i (x_i + u_i)^{k_i} L_i^{1-k_i} \tag{4.7}$$

Donc, la tâche du Centre — l'objectif de la planification — est de trouver une répartition des ressources $U$:

$$U = \sum_{i=1}^{N} u_i, \tag{4.8}$$

qui maximise la fonction (4.1). Le résultat de la répartition des ressources ne dépendra pas seulement des quantités $u_i$, soumises à la condition (4.8), que le Centre désignera, mais aussi de la quantité de travail $L_i$ utilisée par le Producteur. Le Centre sait qu'au moment de faire son choix le Producteur connaîtra la quantité $u_i$, c'est-à-dire le choix du Centre. On conviendra en outre que le Centre connaît la fonction objectif (4.6) du Producteur. L'information dont dispose le Centre lui fait supposer que le Producteur choisira une valeur de $L_i$ qui réalisera le maximum de

$$J_i = c_i \alpha_i (x_i + u_i)^{k_i} L_i^{1-k_i} - \omega_i L_i.$$

Ce problème admet toujours une solution pour $x_i$ et $u_i$ fixes, une solution qui se détermine aisément sous forme explicite à partir de la condition $\partial J_i/\partial L_i = 0$. On trouve

$$L_i = \hat{c}_i (x_i + u_i), \tag{4.9}$$

où

$$\hat{c}_i = \left[ \frac{c_i \alpha_i}{\omega_i} (1 - k_i) \right]^{1/k_i}.$$

Grâce à (4.9), on peut ramener immédiatement le problème de planification à un problème de programmation mathématique. En vertu de (4.9)

$$P_i = \hat{c}_i^{1-k_i} \alpha_i (x_i + u_i) = \beta_i + \gamma_i u_i,$$

c'est-à-dire que le résultat des activités du Producteur est une fonction linéaire des ressources allouées. Donc

$$J(P_1, \ldots, P_N) = J(\beta_1 + \gamma_1 u_1, \ldots, \beta_N + \gamma_N u_N) = \\ = J^*(u_1, \ldots, u_N), \tag{4.10}$$

et l'on est conduit à un problème de maximisation de la fonction (4.10) sous la contrainte linéaire (4.8).

Ainsi, l'hypothèse relative au comportement des Producteurs a permis au Centre de les traiter comme d'ordinaires échelons réflexifs.

REMARQUE. Si la fonction objectif du Centre est linéaire, par exemple

$$J = \sum_{i=1}^{N} P_i, \tag{4.11}$$

sa maximisation est un problème trivial: il faut calculer les coefficients $\gamma_i$ et allouer toutes les ressources au Producteur dont le coefficient $\gamma_i$ est le plus grand.

b) *Gestion par pénalisation et récompense.* Dans l'exemple du numéro a) le Producteur a été mis en possession d'une quantité bien définie de ressources ne dépendant pas de la manière dont elle sera utilisée et des résultats des activités du Producteur. Ce moyen d'action sur le Producteur ne permettait pas au Centre d'influencer la fonction objective de ce dernier.

Il convient maintenant de rappeler un fait sur lequel nous aurons l'occasion de revenir à maintes reprises: l'existence des objectifs personnels des sous-systèmes qui est une réalité objective. Si un élément du système a le droit de se servir des ressources, il se mue inévitablement en organisme dont le but essentiel est la préservation de sa propre homéostasie. Les conditions d'homéostasie varient avec les situations et les organismes. En effet, si la structure des rapports sociaux est telle que la stabilité d'un groupe de production

(d'un organisme) est assurée par le niveau de son revenu, alors la fonction objective du Producteur peut être représentée sous la forme (4.6). Mais il est fort possible par exemple que la condition essentielle qui garantira cette stabilité ne sera pas le revenu, mais une exigence de la direction régionale liée à la non-pollution de l'environnement. Le critère du Producteur sera alors différent. Il dépendra directement de la structure des rapports de production et de l'infrastructure juridique de la société. Le Centre ne peut rien contre ce fait. Il ne peut modifier les rapports sociaux qui, pour mesure de stabilité de l'organisme, préconisent le revenu de cet organisme ou d'autres facteurs. Mais le Centre peut modifier le revenu du Producteur l'obligeant *ipso facto* à agir dans le sens qui lui profite le plus. Les mécanismes économiques de gestion se prévalent justement de ce principe. Expliquons ceci sur l'exemple où le Centre peut peser non pas sur la structure mais sur la valeur de l'objectif (4.6) en la modifiant en fonction des solutions choisies par le Producteur.

Plaçons-nous dans les conditions de l'exemple du n° a) et supposons que l'activité du Producteur est décrite par une fonction de production de Kobb-Douglas (4.4) dans laquelle, pour simplifier, nous poserons tous les $k_i = 1/2$, $i = 1, 2, \ldots, N$. Mettons l'objectif (le revenu) du Producteur sous la forme

$$J_i(L_i) = c_i \alpha_i x_i^{1/2} L_i^{1/2} - \omega_i L_i + \varphi_i(P_i), \qquad (4.12)$$

où $\varphi_i(P_i)$ est une prime (ou une pénalisation) accordée par le Centre au Producteur selon la qualité de ses résultats. Les quantités $x_i$ sont supposées fixes. Ceci permet de mettre la fonction (4.12) sous la forme

$$J_i(L_i) = \hat{\alpha}_i L_i^{1/2} - \omega_i L_i + \varphi_i(P_i), \qquad (4.13)$$

où $\hat{\alpha}_i = c_i \alpha_i x_i^{1/2}$. L'attitude du Producteur n'a pas changé par hypothèse: on admet qu'il choisit sa commande à partir de la condition de maximum de son revenu et par ailleurs le Centre a bien sûr intérêt à ce que le Producteur sache au moment de prendre sa décision les conditions de récompense ou de pénalisation, c'est-à-dire le caractère de la fonction $\varphi_i(P_i)$.

La condition

$$\frac{\partial J_i}{\partial L_i} = \frac{1}{2} \hat{\alpha}_i L_i^{-1/2} - \omega_i + \frac{d\varphi_i}{dP_i} \frac{dP_i}{dL_i} = 0 \qquad (4.14)$$

permet de déterminer la valeur $L_i = L_i^*$ qui réalise le maximum de la fonction $J_i(L_i)$. La quantité $L_i^*$ est une fonctionnelle qui dépend de la forme de la fonction $\varphi_i(P_i)$: $L_i^* = L_i^*[\varphi_i(P_i)]$. De façon analogue, le volume de production optimal $P_i^*$ sera une fonctionnelle: $P_i^* = P_i^*[\varphi_i(P_i)]$, c'est-à-dire qu'il dépendra aussi

de la structure de la fonction $\varphi_i$. Donc la détermination de la commande optimale du Centre se ramène à celle de fonctions de pénalisation et de récompense $\varphi_i (P_i)$ qui maximisent l'objectif du Centre $J (P_1, \ldots, P_N)$. Mettons cette fonction sous la forme

$$J = J (P_1^* [\varphi_1], P_2^* [\varphi_2], \ldots, P_N^* [\varphi_N]). \tag{4.15}$$

La détermination de l'extrémum de la fonctionnelle (4.15) est un problème complexe non classique, car les fonctionnelles $P_i^*$ sont elles-mêmes acquises par la résolution d'un problème d'extrémum très compliqué: $J_i \Rightarrow \max$. Même pour le cas particulier simple du système hiérarchique considéré, la détermination de la pénalisation optimale implique des méthodes spéciales.

En effet, soient $\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_N$ les quantités de biens fabriqués par les Producteurs qui maximisent l'objectif du Centre. Alors la fonction de pénalisation peut être donnée par exemple sous la forme

$$\varphi_i (P_i) = \lambda_i (P_i - \hat{P}_i)^2 - c_i \alpha_i x_i^{1/2} L_i^{1/2} + \omega_i L_i, \tag{4.16}$$

$$i = 1, \ldots, N,$$

où $\lambda_i$ sont des nombres arbitraires $<0$. Dans ce cas l'objectif du Producteur $i$ devient

$$J_i = \lambda_i (P_i - \hat{P}_i)^2.$$

Pour s'assurer un revenu maximal, le Producteur doit user de ses ressources de sorte à fabriquer une quantité de biens $\hat{P}_i$. La structure de la pénalisation fait se confondre les intérêts du Producteur et du Centre.

Il est immédiat de voir que les fonctions de la forme (4.16) qui identifient les intérêts du Producteur et du Centre sous réserve que la fonction $\varphi_i$ ne soit assujettie à aucune contrainte, peuvent être construites d'une infinité de façons. A signaler toutefois le faible intérêt d'une pénalisation ou d'une prime illimitées. Dans les problèmes « assez réels », la pénalisation est ou bien bornée, c'est-à-dire vérifie des conditions de la forme

$$\varphi_i \in G_\varphi, \tag{4.17}$$

où $G_\varphi$ est un ensemble, ou bien la valeur de la fonction $J$ dépend des fonctions $\varphi_i$:

$$J = J [P_1, \ldots, P_N, \varphi_1, \ldots, \varphi_N].$$

Le problème d'optimisation posé présente la même difficulté que les problèmes de synthèse. Il faut chercher des fonctions $\varphi_i (P_i)$ dépendant des coordonnées de phase.

On dispose actuellement de deux approches pour résoudre ces problèmes. L'une d'elles utilise l'idée, traditionnelle en théorie

de la synthèse, de la paramétrisation de la fonction cherchée, l'autre, est reliée à un théorème de Yu. Hermeyer qui affirme l'équivalence entre la recherche d'une solution optimale dans un jeu hiérarchique à deux personnes et un problème spécial de programmation non linéaire (cf. [72]). Etudions ces deux approches pour le cas de deux sujets: le Centre et le Producteur.

Supposons que le Centre dispose du choix de l'élément $x$ et le Producteur de l'élément $y$. Ces deux sujets poursuivent les objectifs respectifs suivants:

$$F(x, y) \Rightarrow \max, \tag{4.18'}$$

$$f(x, y) \Rightarrow \max. \tag{4.18''}$$

La stratégie du Centre est une fonction $x = \psi(y)$. Celle-ci est communiquée au Producteur au moment où il choisit $y$. On admet que le Producteur fait confiance au Centre et sachant $\psi(y)$ détermine $y$ à partir de la condition

$$f(\psi(y), y) \Rightarrow \max. \tag{4.18'''}$$

La résolution de ce problème nous donne un opérateur $y = Y[\psi(\cdot)]$. Il ne reste ensuite au Centre qu'à choisir la fonction $\psi(\cdot)$ à partir de la condition

$$\sup_{\psi(\cdot)} \inf_{y \in Y[\psi(\cdot)]} F(\psi(y), y).$$

En particulier, si le problème (4.18''') admet une solution unique pour tout $\psi(\cdot)$, on est amené à déterminer une fonction $\psi(\cdot)$ réalisant

$$\sup_{\psi(\cdot)} F(\psi(y), y),$$

où $y$ est la solution du problème (4.18''') pour $\psi(\cdot)$ donnée.

La première approche est la paramétrisation de la fonction $\psi(y)$. Elle consiste à représenter $\psi(y)$ comme une fonction de plusieurs paramètres, par exemple

$$\psi(y) = ay + by^2. \tag{4.19}$$

REMARQUE. La détermination de la classe des fonctions de pénalisation et de récompense est un problème spécial ardu. En effet, l'extension de la classe des fonctions de pénalisation et de récompense admissibles peut considérablement modifier la valeur de la fonction objectif. Yu Hermeyer [4] a cité des exemples montrant que l'introduction de fonctions discontinues dans cette classe modifie le résultat final d'autant qu'on le veut.

En se servant de (4.19), on peut mettre le problème (4.18'') sous la forme

$$f^*(a, b, y) \Rightarrow \max,$$

d'où il résulte que si ce problème admet une seule solution pour tous

$a$ et $b$, alors $y$ est une fonction des paramètres $a$ et $b$:

$$y = y(a, b),$$

et le problème (4.18″) se transforme en un problème spécial de programmation mathématique, identique à celui du n° a) de ce paragraphe.

L'autre approche est liée au fait suivant. Considérons le problème

$$f(x, y) \Rightarrow \min_{x}. \qquad (4.20')$$

Sa solution est une fonction $x = x^*(y)$ que l'on conviendra d'appeler « stratégie de pénalisation ». La raison de cette dénomination apparaîtra plus loin. Si le Producteur choisit la stratégie $y$, la fonction

$$\bar{f}(y) = f(x^*(y), y)$$

définit son plus mauvais score.

Pour s'assurer son espérance $f^*$, le Producteur doit résoudre encore un problème d'optimisation:

$$\bar{f}(y) \Rightarrow \max_{y} \qquad (4.20'')$$

dont la solution sera désignée par $f^*$ et la stratégie correspondante par $y^*$.

Considérons maintenant l'objectif $F(x, y)$ du Centre et définissons les solutions $\hat{x}$ et $\hat{y}$ du problème

$$F(x, y) \Rightarrow \max_{x, y}. \qquad (4.20''')$$

On a l'inégalité évidente

$$f(\hat{x}, \hat{y}) \geqslant \bar{f}(y).$$

Mais

$$\bar{f}(y) \leqslant f^*$$

et il se peut que

$$f(\hat{x}, \hat{y}) < f^*.$$

Comme le Producteur est assez indépendant et peut toujours s'assurer le résultat $f^*$, le Centre n'a aucune raison d'espérer que le Producteur optera pour la stratégie $\hat{y}$. Donc, le Centre n'a intérêt à se livrer à une analyse qu'à la condition que son choix vérifie la relation

$$f(x, y) \geqslant f^*. \qquad (4.21)$$

La solution du problème (4.20‴), (4.21) définit des vecteurs $x^0$ et $y^0$. Le théorème de Hermeyer dit que la stratégie optimale du Centre sera une fonction $x(y)$ *)

*) Ce résultat est valable si l'adhérence de l'ensemble $\{x, y : f(x, y) > f^*\}$ est $\{x, y : f(x, y) \geqslant f^*\}$, ce qui a généralement lieu. Pour plus de détails voir [4].

$$x(y) = \begin{cases} \bar{x}^0 & \text{pour } y = \bar{y}^0, \\ x^*(y) & \text{pour } y \neq \bar{y}^0. \end{cases} \tag{4.22}$$

Si $f(x^0, y^0) > f^*$, alors $\bar{x}^0 = x^0$, $\bar{y}^0 = y^0$. Si $f(x^0, y^0) = f^*$, alors $\bar{x}^0$, $\bar{y}^0$ sont tels que $f(\bar{x}^0, \bar{y}^0) > f^*$ et ils sont solutions du problème (4.20‴) à $\varepsilon$ près choisi par le Centre. Si donc l'on résout les problèmes d'optimisation (4.20′), (4.20″) et (4.20‴), alors la fonction de synthèse $x(y)$ s'explicite sous la forme (4.22). Ce fait est remarquable, car il offre d'intéressantes perspectives pour la construction effective des mécanismes de gestion dans les systèmes économiques.

La voie préconisée par le théorème pour la construction des fonctions de pénalisation et de récompense peut paraître assez compliquée. De prime abord elle le paraît plus que la méthode traditionnelle de construction de la fonction de synthèse. Mais, en réalité elle est souvent d'une très grande efficacité pour la construction des mécanismes de gestion dans les systèmes hiérarchiques.

Illustrons ses possibilités par une interprétation économique des raisonnements cités. Considérons le problème (4.20′). Nous pouvons le traiter comme un problème de détermination des actions du Centre qui soient les plus gênantes pour le Producteur. Mais comme ces actions sont toujours soumises à des contraintes et qu'elles sont en nombre limité, le problème (4.20′) est en général trivial. Par exemple, les prix doivent être des prix planchers, la fonction de récompense, égale à zéro, la pénalisation, la plus grande possible, etc.

Le problème (4.20″) est pour le Producteur le problème de sélection de sa meilleure stratégie dans les conditions les plus embarrassantes, la quantité $f^*$, son espérance.

Le problème (4.20‴) est pour le Centre le problème de sélection de la stratégie optimale dans des conditions de centralisation totale mais avec la contrainte (4.21) qui dit que les intérêts du Producteur doivent être pris en considération, c'est-à-dire que son résultat doit être $\geqslant f^*$.

Donc, les problèmes d'optimisation (4.20) sont assez souvent peu compliqués et admettent une interprétation économique simple.

Par ailleurs, la solution $(x^0, y^0)$ peut être traitée comme un programme concerté, puisque le Producteur a intérêt à s'y conformer s'il veut obtenir la plus grande récompense. Le rejet de cette solution concertée est immédiatement pénalisé: la rétroaction entre automatiquement en jeu.

La fonction (4.22) est de la forme représentée sur la figure 3.4. Cette fonction est discontinue. Excepté le point $y = y^0$, elle est partout confondue avec la fonction $x^*(y)$, solution du problème (4.20′) qui définit les plus mauvaises conditions de fonctionnement du Producteur, c'est-à-dire la pénalisation maximale.

La solution (4.22) n'est pas très commode à l'usage. Il se trouve que la fonction (4.22) peut varier dans des limites assez vastes. Dans de nombreuses situations réelles, la stratégie optimale du Centre n'est pas unique et l'on peut remplacer la fonction discontinue (4.22) par une fonction régulière comme l'indique la figure 3.5 *).

Ces remarques faites, revenons à l'exemple considéré. On rappelle

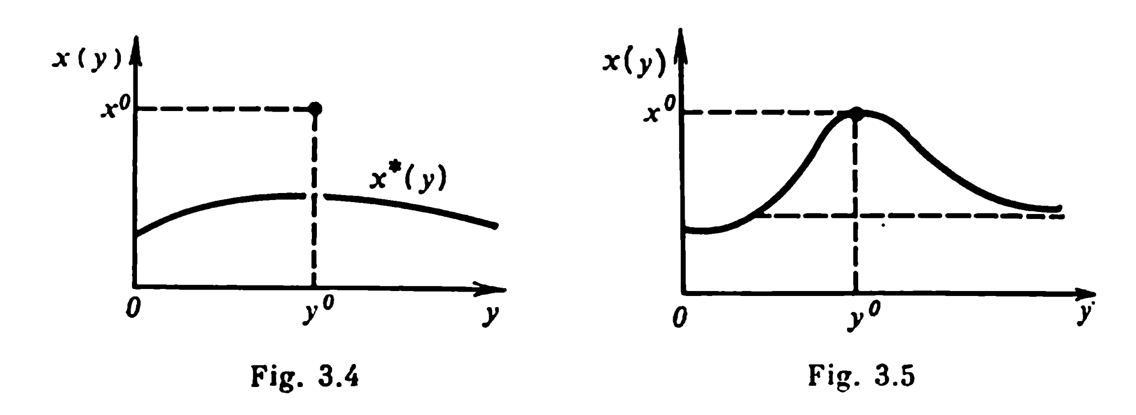



Fig. 3.4 Fig. 3.5

que le Producteur tente de maximiser la fonction (4.13) en agissant sur une seule variable: la quantité de main-d'œuvre $L_i$.

Comme le volume de la production ne dépend que de la main-d'œuvre, on peut chercher la fonction $\varphi_i(P_i)$ comme une fonction de $L_i^{1/2}$ seulement. Désignons $L_i^{1/2}$ par $z_i$.

Le problème du Producteur est donc

$$J_i(z_i) = \hat{\alpha}_i z_i - \omega_i z_i^2 + \varphi_i(z_i) \Rightarrow \max$$

avec les conditions $z_i \geqslant 0$, $\varphi_i(z_i) \geqslant 0$. Résolvons ce problème avec les deux méthodes. Voyons d'abord la première.

Posons $\varphi_i(z_i) = c_{i1}z_i + c_{i2}z_i^2$. La condition $\partial J_i/\partial z_i = 0$ nous donne

$$z_i = \frac{\hat{\alpha}_i + c_{i1}}{2(\omega_i - c_{i2})}.$$

Supposons maintenant que l'objectif du Centre est de la forme

$$J = \sum_{i=1}^{N} k_i z_i - \sum_{i=1}^{N} \varphi_i(z_i) \qquad (4.23)$$

ou

$$J = \sum_{i=1}^{N} k_i z_i - \sum_{i=1}^{N} (c_{i1}z_i + c_{i2}z_i^2) = \sum_{i=1}^{N} \Phi_i,$$

*) Ce chapitre de la théorie des systèmes hiérarchiques est actuellement bien élaboré. Outre l'ouvrage mentionné de Yu. Hermeyer [4] on peut par exemple se référer à [42, 70].

où

$$\Phi_i = (k_i - c_{i1}) z_i - c_{i2} z_i^2.$$

En se servant de l'expression de $z_i$, on obtient

$$\Phi_i = \frac{k_i - c_{i1}}{2(\omega_i - c_{i2})}(\hat{\alpha}_i + c_{i1}) - \frac{c_{i2}(\hat{\alpha}_i + c_{i1})^2}{4(\omega_i - c_{i2})^2} = \Phi_i^*(c_{i1}, c_{i2}).$$

La fonction $\Phi_i^*(c_{i1}, c_{i2})$ est de forme assez simple et sa maximisation ne soulève aucune difficulté.

Voyons maintenant la deuxième méthode de résolution du problème. Comme $\varphi_i(z_i) \geqslant 0$, la plus mauvaise valeur de la fonction de récompense est $\varphi_i = 0$. Alors

$$J_i(z) = \hat{\alpha}_i z_i - \omega_i z_i^2$$

et l'espérance sera

$$J_i^* = \hat{\alpha}_i^2 / 4\omega_i.$$

Le problème (4.20''), où $F$ est de la forme (4.23), s'écrit:

$$\Phi_i = k_i z_i - \varphi_i(z) \Rightarrow \max$$

sous réserve que $J_i \geqslant \hat{\alpha}_i^2/4\omega_i$. De là on déduit immédiatement que

$$\varphi_i = 0, \qquad z_i = \hat{\alpha}_i / 2\omega_i.$$

Dans le cas considéré, la seconde approche est à préférer et pas seulement parce qu'elle nous conduit plus rapidement au résultat final: la résolution du problème par paramétrisation est aussi assez simple. La raison est ailleurs. Dans le premier cas, nous n'avons pas réussi à dégager un résultat pourtant évident: la coïncidence des intérêts du Centre et du Producteur (le Producteur choisira $L_i = \hat{L}_i$ (fig. 3.6)). La paie de la main-d'œuvre pour $L_i > \hat{L}_i$ n'est pas amortie par le prix de revient du bien produit. Mais ce choix ne profite pas non plus au Centre, puisque ce dernier verse à la main-d'œuvre supplémentaire par le biais de la fonction de récompense la même somme que le Producteur.

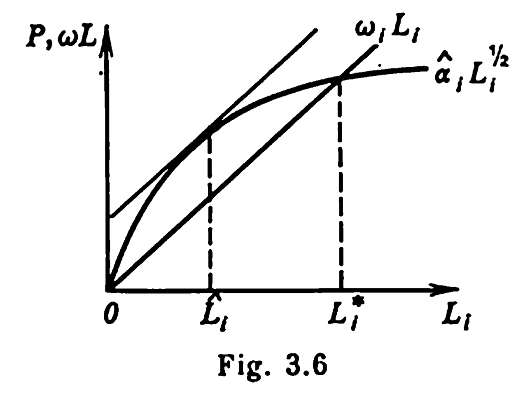


Fig. 3.6

c) *Utilisation des ressources exogènes.* Voyons encore un problème formellement équivalent à celui que nous venons tout juste d'examiner. Au nº a) nous avons étudié une méthode de gestion d'entreprises à l'aide de leurs ressources initiales sans les rattacher directement au résultat de leur utilisation, c'est-à-dire au volume de la production.

Revenons maintenant au modèle de distribution des ressources, mais convenons que les ressources initiales sont représentées par une fonction des biens produits $P_i$:

$$u_i = u_i (P_i),$$

et supposons que les ressources allouées par le Centre sont entièrement investies dans la construction de nouvelles unités.

Nous conservons l'ancienne hypothèse sur le comportement du Producteur, c'est-à-dire que nous admettrons que le Producteur choisira sa stratégie à partir de la condition de maximisation de son revenu $J_i$:

$$J_i = c_i \alpha_i (x_i + u_i (P_i))^{k_i} L_i^{1-k_i} - \omega_i L_i.$$

Supposons que la fonction $u_i (P_i)$ est donnée. Vu que nous n'imposons aucune contrainte sur $L_i$, la quantité $L_i$ sera déduite à partir de la condition

$$\frac{\partial J_i}{\partial L_i} \equiv c_i \frac{\partial P_i}{\partial L_i} - \omega_i = 0. \tag{4.24}$$

Calculons la dérivée $\partial P_i/\partial L_i$ en traitant la relation

$$P_i - \alpha_i (x_i + u_i (P_i))^{k_i} L_i^{1-k_i} = 0$$

comme une fonction implicite $P_i$ de $L_i$, d'où

$$\left[1 - k_i \alpha_i (x_i + u_i (P_i))^{k_i - 1} L_i^{1-k_i} \frac{du_i}{dP_i}\right] \frac{\partial P_i}{\partial L_i} - \\ - (1 - k_i) L_i^{-k_i} \alpha_i (x_i + u_i (P_i))^{k_i} = 0.$$

De là on déduit

$$\frac{\partial J_i}{\partial L_i} = c_i \frac{(1 - k_i) \alpha_i (x_i + u_i (P_i))^{k_i} L_i^{-k_i}}{1 - k_i \alpha_i (x_i + u_i (P_i))^{k_i - 1} L_i^{1-k_i} \frac{du_i}{dP_i}} - \omega_i = 0. \tag{4.25}$$

Donc, dans le cas considéré, le problème du Centre se ramène au suivant: déterminer des fonctions positives $u_i (P_i)$ réalisant le maximum de la fonctionnelle $J (P_1, \ldots, P_N)$ sous les contraintes (4.25).

On pourrait citer beaucoup d'autres exemples illustrant les possibilités de gestion dans les systèmes hiérarchiques. Par exemple, une possibilité qui est presque toujours utilisée dans n'importe quel système hiérarchique est la restriction de l'activité des échelons inférieurs. Citons un exemple classique: la limitation des fonds des salaires:

$$\omega_i L_i \leqslant Q_i, \qquad i = 1, \ldots, N. \tag{4.26}$$

Jusqu'à maintenant nous avons admis que la fonction de production $P_i = f_i(x_i, L_i)$ de l'entreprise $i$ ne dépendait que des fonds $x_i$ et de la main-d'œuvre $L_i$. En réalité, le taux des salaires qui modifie les conditions d'embauche de la main-d'œuvre joue aussi un grand rôle. En d'autres termes, $P_i = f_i(x_i, L_i, \omega_i)$ et le taux des salaires doit naturellement être considéré comme un facteur de gestion dépendant du Producteur. Mais la quantité $Q_i$ doit alors être incluse dans le nombre de facteurs gouvernés par le Centre. La condition (4.26) nous donne encore un procédé de gestion.

REMARQUE. Signalons que la méthode de gestion formellement décrite est parfaitement identique (du point de vue mathématique) à la méthode de gestion par répartition des ressources. Mais en réalité toutes les méthodes de gestion décrites présentent tellement de particularités non formelles qu'il est nécessaire de les étudier séparément. Les moyens d'action de l'échelon supérieur sur les échelons inférieurs, c'est-à-dire les moyens d'action sur leurs intérêts et les moyens de détournement de ces intérêts pour la réalisation de certains objectifs de l'échelon supérieur, se définissent en principe au terme d'une analyse non formelle. Quant au rôle du mathématicien, il consiste à éclaircir les aspects quantitatifs, à dégager des relations cohérentes entre les paramètres de gestion et les états du système.

d) *Sur une décentralisation rationnelle et l'efficacité des structures décentralisées.* Le problème d'une décentralisation rationnelle est un problème clé en économie. Si un organisme économique devient assez compliqué, sa décentralisation totale, ainsi qu'on l'a signalé plus haut, peut facilement conduire à des solutions erronées en raison de l'impossibilité de traiter à temps l'information nécessaire. Mais une décentralisation trop poussée est aussi lourde de conséquences: certains échelons se transforment en organismes indépendants et commencent à poursuivre des objectifs personnels. Dans ces conditions, il n'est pas exclu que l'avantage d'une gestion décentralisée apte à tenir pleinement compte des particularités de l'information soit réduit à néant par une intense activité des échelons qui sont détournés des intérêts du système. Voilà pourquoi la liberté d'action accordée par l'échelon supérieur à ses subordonnés doit faire l'objet d'une analyse spéciale. Sa définition est un problème clé de la théorie des systèmes décentralisés de gestion. On observera que les positions des problèmes de cette théorie se distinguent foncièrement de ceux rencontrés dans le cadre de la théorie traditionnelle des deux antagonistes.

Soit donné un système à deux sujets $X$ et $Y$. Les objectifs des sujets $X$ et $Y$ sont respectivement

$$F(x, y) \Rightarrow \max_{x \in G_x},$$

$$F(x, y) \Rightarrow \min_{y \in G_y}.$$

Supposons que le sujet $X$ a la possibilité de modifier l'ensemble des

stratégies $G_y$ du sujet $Y$. Il est évident qu'il aura toujours intérêt à rétrécir l'ensemble des stratégies de son partenaire (son adversaire dans ce cas, puisque la situation envisagée est antagoniste).

La situation est différente en théorie des systèmes hiérarchiques de gestion. Les sujets de ce système possèdent des objectifs personnels qui ne sont nullement en contradiction avec ceux du Centre. Donc, le Centre aura peut-être intérêt à élargir les droits de certains échelons. Mais cet élargissement des droits (c'est-à-dire l'extension de l'ensemble des stratégies permises) n'est profitable que dans une certaine mesure. La liberté d'action accordée aux échelons inférieurs dépend évidemment des particularités de chaque système et varie d'un système à l'autre. Il n'existe pas de recettes universelles en la matière.

Ainsi, projeter un système hiérarchique de gestion consiste tout d'abord à construire un modèle de fonctionnement du système pour une structure donnée. S'il s'agit d'un système hiérarchique en éventail, cette structure sera définie par l'énumération des échelons indépendants. On rappelle que dans ces conditions chaque échelon devient un organisme autonome. Donc, l'étape suivante consiste à dégager les objectifs des échelons inférieurs. Il serait plus logique de parler non pas d'objectifs mais d'intérêts, bien que ces deux notions soient identiques formellement. Certes, il est pratiquement impossible d'indiquer exactement l'objectif de l'échelon inférieur (du Producteur): les intérêts du Producteur sont toujours décrits par un système de critères, et le choix de la réduction des critères dont la maximisation détermine le comportement du Producteur est toujours teinté de subjectivisme. Cependant pour que le Centre puisse définir une politique, il est nécessaire qu'il pose une condition sur les objectifs des Producteurs (condition de comportement). Après avoir « défini » les objectifs des Producteurs, nous devons élucider les moyens de gestion du Centre et des Producteurs. Enfin, en dernier ressort nous pouvons réduire le problème à un problème spécial d'optimisation. Ces problèmes seront identiques à ceux décrits dans ce paragraphe. L'analyse de ces problèmes nous permet de trouver les mécanismes optimaux de fonctionnement (dans le cadre des hypothèses admises): règles de répartition des ressources, système de pénalisation et de primes, degré de la liberté d'action, etc.

Signalons qu'à ce stade de l'analyse, il y a lieu d'introduire la notion de commande optimale d'un système hiérarchique comme une commande réalisant le maximum de l'objectif du Centre sous réserve que les commandes des Producteurs soient déterminées à partir de leurs conditions de comportement. La dernière étape d'élaboration d'une structure hiérarchique consiste précisément à calculer une commande optimale. Quand on projette un nouveau système quel qu'il soit, on veut toujours s'assurer qu'il sera mieux

que l'ancien et de plus on essaye de l'apprécier au niveau déjà du projet, bien avant sa fabrication et *a fortiori* avant son fonctionnement. Il est entendu que ceci vaut également pour l'élaboration des systèmes hiérarchiques de gestion.

Signalons que les raisonnements utilisés dans ce paragraphe ne nous permettent pas en principe de procéder à une comparaison de l'efficacité des divers systèmes de gestion, car nous avons étudié le fonctionnement d'un système hiérarchique dans des conditions d'information complète. Or toute décentralisation, c'est-à-dire la substitution d'un système hiérarchique à un système entièrement centralisé, n'est pas rentable dans le cadre de cette théorie.

Soit $J^*$ la valeur optimale de l'objectif du Centre dans un système de gestion hiérarchique. Supposons maintenant que le système est entièrement centralisé, c'est-à-dire que le Centre peut non seulement répartir les ressources exogènes, fixer les pénalisations, etc., mais aussi définir la quantité de main-d'œuvre. Dans ces conditions appelons $J^{**}$ le maximum de l'objectif. Il est alors évident que

$$\Delta = \frac{J^{**} - J^*}{J^*} \geqslant 0$$

puisque la hiérarchisation du système n'est qu'une contrainte supplémentaire restreignant l'ensemble des stratégies admissibles. Donc, pour apprécier si une hiérarchisation est adéquate, il faut nécessairement tenir compte de la modification du caractère de l'information qu'elle entraîne. Le but de la hiérarchisation est tout d'abord de diminuer le niveau d'indétermination dans les procédures de prise de décision. Nous devons donc nécessairement retourner aux raisonnements du début du paragraphe précédent. Reprenons ces raisonnements pour le cas d'une répartition des ressources exogènes dans un système hiérarchique à deux échelons.

La fonction de production du Producteur sera prise de la forme

$$P_i = (\alpha_i + \xi_i)\, x_i^{k_i} L_i^{1-k_i}, \quad i = 1, \ldots, N,$$

où $\xi_i$ est un paramètre connu exactement du Producteur mais inconnu au Centre. La quantité $\alpha_i$ caractérise l'efficacité des fonds, elle dépend de la situation de l'entreprise considérée (qualité des locaux, qualification de la main-d'œuvre, niveau d'organisation, etc.).

Etant donné que le Producteur connaît exactement ses possibilités, en fixant la quantité de main-d'œuvre, il connaîtra avec précision le volume de la production finale. La situation est différente pour le Centre qui est moins renseigné. Supposons par exemple que le Centre ne connaît que l'intervalle de variation de $\xi$:

$$\xi_i^- \leqslant \xi_i \leqslant \xi_i^+.$$

Supposons que le Centre a opté pour une centralisation totale, c'est-à-dire que non seulement il répartit les ressources $u_i$ entre les

Producteurs, mais aussi a le droit de désigner les quantités $L_i$. Son objectif sera alors de la forme

$$J = J(P_1, \ldots, P_N) =$$

$$= J((\alpha_1 + \xi_1)(x_1 + u_1)^{k_1} L_1^{1-k_1}, \ldots, (\alpha_N + \xi_N)(x_N + u_N)^{k_N} L_N^{1-k_N}),$$

où (4.27)

$$J = \tilde{J}(u_1, \ldots, u_N, L_1, \ldots, L_N, \xi_1, \ldots, \xi_N).$$

Avant de chercher le maximum de la fonction (4.27), nous devons soumettre le critère $J$ à une certaine condition. La caractéristique la plus naturelle du système sera l'espérance maximale

$$I^{**} = \max_{u_i, \sum u_i = U} \max_{L_i} \min_{\xi_i \in [\xi_i^-, \xi_i^+]} \tilde{J}(u_i, \ldots, u_N, L_1, \ldots, L_N, \xi_1, \ldots, \xi_N).$$

Considérons maintenant le cas où le système est muni d'une structure hiérarchique de prise des décisions. Comment apprécier la valeur de la fonctionnelle du Centre? Nous pouvons construire l'espérance. Nous devons à cet effet résoudre le problème

$$J_i = J_i(L_i, u_i, \xi_i) \Rightarrow \max. \qquad (4.28)$$

La résolution du problème (4.28) nous donne

$$L_i = L_i(u_i, \xi_i);$$

la fonctionnelle (4.27) peut alors être mise sous la forme

$$J = \hat{J}(u_1, \ldots, u_N, \xi_1, \ldots, \xi_N),$$

et par suite

$$I^* = \max_{u_i, \sum u_i = U} \min_{\xi_i \in [\xi_i^{-1}, \xi_i^+]} \hat{J}(u_1, \ldots, u_N, \xi_1, \ldots, \xi_N).$$

La quantité trouvée est aussi une estimation importante et utile de la structure hiérarchique.

Pour les fonctionnelles $I^*$ et $I^{**}$ nous pouvons de nouveau composer les quantités

$$\Delta = \frac{I^{**} - I^*}{I^*}.$$

Dorénavant nous ne pouvons plus affirmer que $\Delta$ est toujours $\geqslant 0$. Mais cette caractéristique est-elle suffisante pour justifier la hiérarchisation du système? Certainement pas dans le cas général.

Dans la réalité, d'autres caractéristiques tels la complexité et la fiabilité du système de transmission des données, le coût du traitement de l'information, etc., joueront un rôle important. Pour être définitivement fixé sur l'adéquation du changement d'un système entièrement centralisé par un système hiérarchique, on peut

faire appel avec bonheur à la simulation sur ordinateur pour diverses valeurs de $\xi_i$.

La simulation d'un système sur machine par la méthode de Monte-Carlo peut constituer un très important élément dans l'appréciation de la qualité d'une commande: le projeteur ou le futur usager ont par ces expériences une vision concrète des particularités de fonctionnement du système dans les diverses conditions d'information.

L'indétermination dépend beaucoup des ressources techniques mises en œuvre. L'accroissement de la rapidité des calculatrices, le perfectionnement des terminaux, des systèmes de transmission des données, etc., contribueront probablement à élever le degré de centralisation des systèmes.

Signalons en conclusion que pour réaliser son objectif le Centre n'a pas besoin de connaître ceux des divers éléments du système, il lui suffit seulement de connaître la réponse de ces éléments à son action. Cette circonstance peut entraîner diverses simplifications.

REMARQUE. Nous ne disons rien de la prise en considération des facteurs non formels lors du choix d'une structure hiérarchique. Dans le même temps, l'efficacité d'un système dépend beaucoup de divers facteurs psychologiques et sociaux qui sont tributaires à leur tour de la structure du système, de son organisation. Ainsi, par exemple, dans le contexte d'une décentralisation, les responsables des échelons inférieurs ont de bien plus grandes possibilités d'initiative et de créativité, ce qui à l'évidence contribue à améliorer la qualité du système dans son ensemble.

e) *Dynamique des systèmes hiérarchiques.* Passons maintenant à l'étude des systèmes cybernétiques non stationnaires en nous bornant comme précédemment à l'étude d'exemples.

Considérons de nouveau un système en éventail (une association ou une firme) constitué de $N$ Producteurs (entreprises) produisant des biens $P_1, \ldots, P_N$. Désignons par $x_i$ les fonds fixes du Producteur $i$. Les variations des fonds sont régies par les équations

$$\dot{x}_i = -k_i x_i + u_i(t) + v_i(t), \qquad i = 1, \ldots, N, \qquad (4.29)$$

où $k_i$ est le coefficient d'amortissement, $u_i$, les investissements du Centre (de la firme), $v_i$, les investissements internes ($u_i$ et $v_i$, les commandes respectives du Centre et du Producteur, représentent les investissements de capitaux par unité de temps).

La production sera décrite par des fonctions de production de la forme

$$P_i = \varphi_i(x_i, L_i, \omega_i), \qquad i = 1, \ldots, N, \qquad (4.30)$$

où $L_i$ est la quantité de main-d'œuvre, $\omega_i$, le taux des salaires. Les quantités $L_i$ et $\omega_i$ sont choisies par le Producteur compte tenu des contraintes

$$L_i \geqslant L_i^- > 0, \quad \omega_i \geqslant \omega_i^- > 0, \quad \omega_i L_i \leqslant Q_i, \quad i = 1, \ldots, N.$$

La signification des deux premières contraintes est évidente: pour pouvoir tourner, l'entreprise a besoin d'un minimum de main-d'œuvre et le taux des salaires doit être supérieur à un certain minimum. La dernière inégalité dit que le fonds des salaires est limité. La quantité $Q_i$ est choisie par le Centre : c'est l'une de ses commandes, les $L_i^-$ et $\omega_i^-$ sont supposés donnés.

Après avoir fabriqué un bien décrit par le vecteur $P_i$, le Producteur l'écoule: il l'envoie au dépôt de la firme ou le met sur le marché. Désignons par $c_i$ le vecteur des prix ; le Producteur perçoit la somme $(c_i, P_i)$ par unité de temps. De cette somme il doit déduire la paie $\omega_i L_i$ des ouvriers, la contribution aux fonds de la firme $\gamma_i (P_i)$, les investissements intérieurs $v_i$ et les frais courants $R_i (P_i)$. Désignons par $\Psi_i$ la somme restante:

$$\Psi_i (t) = (c_i, P_i) - [\omega_i L_i + \gamma_i (P_i) + v_i + R_i (P_i)]. \qquad (4.31)$$

Appelons la quantité $\Psi_i (t)$ fonds social de l'entreprise. Cette quantité se trouve à la disposition du Producteur et peut être utilisée pour primer les ouvriers, pour les besoins sociaux, etc. De par sa signification cette quantité est positive pour tout $t \in [0, T]$.

On admettra que les conditions d'homéostasie sont liées au volume du fonds social pour des contraintes de la forme $P_i \geqslant P_i^-$, $i = 1, \ldots, N$. Si donc les investissements extérieurs $u_i (t)$ sont donnés, si est connue la fonction $\gamma_i (P_i)$ qu'il est naturel d'appeler fonction de récompense ou de pénalisation, et si enfin est connu le fonds des salaires $Q_i$, alors le problème du Producteur est d'utiliser les investissements internes $v_i (t)$, les taux des salaires $\omega_i (t)$ et la quantité de main-d'œuvre $L_i (t)$ de sorte à maximiser dans un certain sens son fonds social. Désignons cette fonctionnelle par

$$J_i = J_i (v_i, \omega_i, L_i).$$

Cette fonctionnelle peut être de la nature la plus diverse. On peut par exemple poser

$$J_i = \min_{t \in [0, T]} \Psi_i (t). \qquad (4.32)$$

Maximiser cette fonctionnelle, c'est maximiser le montant minimal du fonds social créé par unité de temps. Il est aussi logique de considérer des fonctionnelles intégrales de la forme

$$J_i = \int_0^T \Psi_i (t)\, dt, \qquad (4.33)$$

où $T$ est l'horizon planifié. Les problèmes faisant intervenir des fonctionnelles (4.32), (4.33) sont des problèmes classiques de commande optimale. Signalons que la quantité $T$ est une caractéristique subjective du Producteur définie aussi bien par des facteurs exogènes

que par ses qualités personnelles (en particulier, par sa clairvoyance).

Voyons maintenant le fonctionnement du conseil d'administration de la firme (du Centre) contrôlant le Producteur. Cette organisation est par essence gestionnaire et ne produit aucun bien. Elle est jugée sur les résultats du fonctionnement des entreprises qu'elle contrôle. Le critère de la firme est de la forme

$$J = J(P_1, \ldots, P_N, \gamma_1, \ldots, \gamma_N). \tag{4.34}$$

L'expression (4.34) indique que le revenu de la firme dépend de la structure des fonctions de récompense $\gamma_i$. Si l'on admet que le critère du Centre est indépendant de ces fonctions et que le fonds des primes est illimité, alors l'assertion suivante est triviale: il est toujours possible de fixer des primes (ou des pénalisations) $\gamma_i(P_i), \ldots$ $\ldots, \gamma_N(P_N)$ qui contraindront le Producteur à choisir ses commandes comme si la gestion de la firme était entièrement centralisée.

Pour fonctionnelle $J$ on peut prendre les quantités les plus diverses: la précision de réalisation d'un plan, la maximisation du nombre d'équipements, le revenu net, etc. Le problème du conseil de la firme consiste à répartir les ressources

$$\sum u_i = U,$$

le fonds des salaires $Q$

$$\sum Q_i = Q$$

et à désigner des fonctions de récompense et de pénalisation $\gamma_i(P_i)$ de sorte à maximiser le revenu (4.34).

La firme dispose donc de trois moyens d'action sur le Producteur: la répartition des ressources, l'introduction des fonctions de récompense et de pénalisation $\gamma_i(P_i)$ et la limitation de l'activité des Producteurs.

Cette situation conduit selon la terminologie en usage à un jeu différentiel à $N + 1$ personnes. Il n'existe pas de théorie générale qui permette d'étudier cette situation à partir d'un principe plus ou moins unique. Les problèmes classiques de R. Isaacs, L. Pontriaguine, N. Krassovski et autres sont des problèmes de jeux antagonistes et les résultats acquis par ces derniers sont peu utiles à l'analyse des problèmes examinés dans cet ouvrage (cf. [36], [47]).

A. Kononenko a réussi à appliquer les idées de Yu. Hermeyer aux jeux différentiels non antagonistes et à développer sur la base de ces idées un formalisme assez général (cf. [43], [44]) qui laisse entrevoir la possibilité de créer des procédures de calcul efficaces. D'autre part, les recherches de A. Kononenko permettent de considérer sous un angle nouveau la situation prévalant en théorie des jeux différentiels. Les résultats acquis en théorie des jeux non antagonistes ne peuvent être transposés de façon continue à la théorie des jeux antagonistes. Ces derniers sont une dégénérescence singulière

et il leur correspond probablement une construction qui reflète assez mal la réalité: une analyse assez approfondie de certains conflits montre que les intérêts de leurs protagonistes ne sont pas strictement opposés.

L'analyse de tout système hiérarchique débute par la formulation d'une hypothèse sur le comportement de certains de ses échelons. Supposons que cette hypothèse se traduise par la maximisation de la fonctionnelle (4.33).

REMARQUE. Signalons qu'à la différence du cas statique, il nous faut poser des conditions non seulement sur la structure des fonctions $\Psi_i$, mais aussi sur la période de la prévision, c'est-à-dire sur l'horizon $T$ de la planification. Ces quantités varient d'un sujet à l'autre.

Si les actions extérieures du Centre sont fixes, la maximisation de l'objectif du Producteur est un problème de commande optimale et les commandes que choisira le Producteur varieront naturellement avec les actions du Centre. Cela veut dire que le problème du Producteur ne sera pas un problème de commande optimale, mais un problème de synthèse: déterminer les commandes comme des fonctions ou des fonctionnelles des commandes du Centre.

Supposons néanmoins que nous ayons réussi à résoudre ce problème (c'est-à-dire à construire un algorithme efficace). Reste le dernier pas. Nous devons résoudre le problème du Centre: déterminer les fonctions $U_i(t)$, $Q_i(t)$ et $\gamma_i(P_i)$ qui réalisent le maximum de la fonctionnelle (4.34). C'est encore un problème de synthèse mais bien plus épineux que celui du Producteur.

REMARQUE. En effet, le Producteur est confronté dans le cadre de ses activités à un problème bien plus simple que celui du Centre: le Producteur connaît toujours les « conditions du jeu », c'est-à-dire les fonctions et les contraintes retenues par le Centre, et résout le problème pour des valeurs concrètes de ces quantités. Le Centre, pour sa part, doit pour résoudre ses problèmes connaître tous ceux du Producteur, c'est-à-dire pour toutes les valeurs des quantités se trouvant à la disposition du Centre.

Donc, même si l'on a réussi à décomposer l'analyse de l'activité de la firme en une séquence de problèmes d'optimisation (ce qui constitue un progrès indéniable), il est peu probable que l'on puisse proposer des méthodes numériques de résolution assez universelles.

Dans ces problèmes on ne peut visiblement pas se passer des techniques de la simulation et le paragraphe 6 de ce chapitre sera consacré à l'élaboration et l'utilisation des systèmes de simulation. Faisons les observations suivantes.

L'analyse réelle et l'élaboration des systèmes hiérarchiques dynamiques se ramèneront apparemment à la séquence suivante de procédures homme-machine non formelles.

1. L'expert (le projeteur, le gérant) propose un schéma d'organisation, une certaine variante de structure hiérarchique.

2. L'expert détermine avec le concours de l'analyste une variante de la politique du Centre.

3. L'analyste résout les problèmes de commande optimale, trouve la réponse du Producteur, calcule les fonctionnelles du Centre et présente cette information à l'expert.

4. L'expert et l'analyste élaborent une nouvelle politique du Centre, etc.

A la dernière étape, il faut de toute évidence disposer d'un algorithme spécial, puisque le nombre de variantes possibles, même en discrétisant le problème, est très élevé.

Ces procédures homme-machine sont encore intéressantes par le fait que dans des problèmes réels de cette nature il existe toujours pas mal de facteurs non formels dont on peut tenir compte seulement lorsque l'expert intervient dans le processus de calcul de la solution optimale.

Cette procédure rappelle la procédure de Brown-Robinson de résolution des problèmes de la théorie des jeux.

Il existe enfin des méthodes utilisant à des fins de synthèse les idées de Yu. Hermeyer. Donc, la remarque sur l'impossibilité de composer aujourd'hui des algorithmes efficaces de résolution des problèmes considérés relève d'un défaitisme injustifié. La vérité c'est qu'à ce jour aucune tentative sérieuse n'a été entreprise pour l'élaboration de procédures numériques en théorie de la gestion de systèmes hiérarchiques.

f) *Quelques traits particuliers des problèmes récursifs.* Dans les problèmes dynamiques de prise de décision, il existe, contrairement aux problèmes statiques et sous réserve que le système soit oligopole, de nombreux procédés pour lever les indéterminations. En s'imaginant dans les grandes lignes les intérêts et objectifs des échelons inférieurs, le Centre peut au départ ne pas connaître exactement de nombreuses particularités concrètes du fonctionnement des Producteurs. L'observation des activités des échelons inférieurs fournit au Centre une certaine information qui lui permet pour un nombre suffisant de procédures itératives de prise de décision de lever l'indétermination existante et de compléter le tableau. Ce problème s'appelle *problème de commande adaptative*. Ce problème date des années cinquante et l'adaptation est utilisée aujourd'hui avec succès en théorie des systèmes hiérarchiques (cf. [55, 71]). La théorie développée est d'une grande importance pratique. Elle permet en particulier de dévoiler les réserves latentes des Producteurs, des réserves qui une fois connues de ces derniers ne sont pas généralement communiquées aux échelons supérieurs de la hiérarchie.

Eclaircissons certaines des particularités mentionnées sur l'exemple d'un système à temps discret. Supposons que l'objectif du Producteur est de la forme

$$J_i = c_i P_i - \omega_i L_i. \tag{4.35}$$

où $P_i = \alpha_i (x_i + u_i)^{k_i} L_i^{1-k_i}$, $i = 1, \ldots, N$. Ici $\alpha_i$ représentent l'efficacité des fonds, $x_i$, le volume des fonds, $u_i$, les investissements du Centre. Supposons que l'efficacité des fonds est inconnue du Centre. Le problème du Centre est de répartir les investissements entre les Producteurs de manière à maximiser la fonctionnelle additive

$$J = \sum_{i=1}^{N} \sum_{t=0}^{T} c_i P_i(t). \tag{4.36}$$

Au premier pas du processus, le Centre peut répartir les investissements $u_i$ (1) en partant de valeurs inexactes des coefficients d'efficacité $\alpha_i$, par exemple en faisant $\alpha_i = \alpha_{i0}$.

En admettant que le Producteur maximise son revenu à chaque pas (par exemple au cours de chaque année), on trouve à partir de la condition

$$\frac{\partial J_i}{\partial L_i} = (1-k_i)\, c_i \alpha_i\, (x_i + u_i)^{k_i} L_i^{-k_i} - \omega_i = 0,$$

que

$$L_i = \left[\frac{c_i \alpha_i (1-k_i)}{\omega_i}\right]^{1/k_i} (x_i + u_i),$$

où $u_i$ est supposée connue (on la calcule par exemple à partir des conditions posées sur l'efficacité des fonds). Sachant $L_i$, on peut ensuite calculer $P_i$:

$$P_i(1) = \alpha_i (x_i + u_i(1)) \left[\frac{c_i \alpha_i (1-k_i)}{\omega_i}\right]^{(1-k_i)/k_i}. \tag{4.37}$$

Mais la quantité $P_i$ sera connue du Centre autrement que par les calculs: $P_i$ (1) est la quantité de biens produite par le Producteur et envoyée dans les dépôts du Centre. Donc, la formule (4.37) lui permettra de déterminer la quantité inconnue $\alpha_i$.

La signification de la commande adaptative est donc évidente: on avance des hypothèses sur les éventuelles valeurs du paramètre inconnu. Les observations des diverses caractéristiques du processus nous fournissent une nouvelle information qui nous permet de trouver la valeur exacte du paramètre inconnu.

Les raisonnements présumés peuvent être appliqués à l'analyse de cas plus compliqués.

Il se peut par exemple que le Centre ignore deux paramètres: l'efficacité des fonds et, disons, les quantités $x_i$. La relation (4.37) reliera alors deux quantités inconnues: $\alpha_i$ et $x_i$. Au pas suivant on peut calculer

$$P_i(2) = \alpha_i (x_i + u_i(2)) \left[\frac{c_i \alpha_i (1-k_i)}{\omega_i}\right]^{(1-k_i)/k_i} \tag{4.38}$$

et en comparant cette valeur avec la valeur $P_i(2)$ observée, on obtient une deuxième équation pour la détermination de la deuxième inconnue. Mais pour cela il faut se donner $u_i(2)$. Si au premier pas on s'est donné $u_i(1)$ en s'appuyant sur telle ou telle hypothèse sur les valeurs des inconnues $\alpha_i$ et $x_i$, au pas suivant on peut déjà utiliser des raisonnements plus fins eu égard à l'addivité de la fonctionnelle.

La quantité $u_i(2)$ peut par exemple être calculée à partir de la condition de maximum

$$\sum_{i=1}^{N} c_i P_i(2)$$

avec les conditions subsidiaires (4.37) et (4.38), où $u_i(1)$ est supposée connue. On obtient en définitive $\alpha_i$ et $x_i$.

Le schéma de la commande adaptative peut être compliqué davantage si l'on recherche $u_i(1)$ et $u_i(2)$ en maximisant

$$\sum_{i=1}^{N} c_i [P_i(1) + P_i(2)]$$

sous les conditions (4.37) et (4.38).

Le schéma de commande adaptative est un mécanisme très souple qui peut être adapté à l'analyse d'un large éventail de problèmes de prise de décision dans des conditions d'indétermination. La répétition des procédures de prise de décision est susceptible d'apporter à l'analyste l'information qui lui fait défaut. Les fonctions de production peuvent varier au cours de l'activité des entreprises. Cela signifie que leurs paramètres varieront dans le cadre des diverses approximations. L'écart entre les quantités calculées à l'avance et les données des observations indique précisément que les valeurs des paramètres de la fonction de production ont changé. Par des raisonnements analogues à ceux qui viennent d'être effectués, on peut restituer les valeurs inconnues de ces paramètres.

L'utilisation efficace des idées de l'adaptation dépend pour beaucoup de la structure des fonctionnelles considérées. Ces idées permettent assez facilement de restaurer les valeurs des paramètres inconnus si la fonctionnelle du Centre est de la forme additive (4.36). On pourrait développer des procédures plus ou moins efficaces du genre programmation dynamique pour le cas général de fonctions additives. La réalisation d'algorithmes adaptatifs identiques est particulièrement compliquée pour des fonctions objectifs de forme plus générale. Autant que nous sachions il n'existe pas encore d'algorithmes pour le cas général.

12—01523

## § 5. Méthode de programmation dans les systèmes non réflectifs

L'analyse des systèmes non réflectifs et *a fortiori* l'élaboration des commandes dans de tels systèmes est un problème si compliqué qu'on imagine mal comment on pourrait lui trouver une méthode universelle de résolution. Et pourtant vers les années soixante on voit émerger une approche générale pour l'élaboration des principes de gestion dans les systèmes non réflectifs complexes, à laquelle on a donné le nom de méthode de programmation. A strictement parler, ce n'est pas une méthode mais plutôt un ensemble de recommandations assez bien élaborées qui prend les traits d'une méthode seulement dans certains cas spéciaux de gestion de systèmes économiques. Dans ce paragraphe nous décrivons le schéma général de la méthode de programmation dans des systèmes cybernétiques et nous l'appliquerons ensuite à un problème de gestion d'un organisme économique centralisé de type socialiste.

a) *Retour au problème d'optimisation en deux étapes.* Dans les chapitres précédents, nous avons développé déjà les approches d'analyse de problèmes complexes à plusieurs critères et avons notamment mis l'accent sur la nécessité de combiner les méthodes formelles et non formelles pour étudier les problèmes de gestion. Ces idées seront maintenant concrétisées et appliquées à l'élaboration d'approches générales dans les problèmes de choix des commandes dans les systèmes cybernétiques.

Considérons de nouveau un système monopole (un système dynamique commandé)

$$\dot{x} = f(x, u, \xi, t), \tag{5.1}$$

où $\xi(t)$ décrit des actions extérieures aléatoires. Formulons l'objectif de la commande pour ce système. Supposons que la formulation de l'objectif est un acte exogène (par exemple est une doctrine). L'objectif de la commande peut être formulé en termes d'appartenance des trajectoires du système à un certain ensemble, sous forme de conditions imposées aux valeurs aux bornes des variables de phase; l'objectif de la commande peut être un système d'indices de contrôle en économie ou une orbite d'un engin spatial.

L'objectif de la commande est donc représenté par des contraintes. D'autre part, nous formulons des critères qui évaluent les moyens de réalisation de l'objectif de la commande (par exemple les dépenses de carburant ou d'une autre matière), c'est-à-dire que les critères évaluent le coût de la réalisation de l'objectif dans les termes choisis.

Ainsi, la première étape du processus de commande consiste en la définition extrinsèque de l'objectif de la commande et en la formation, en vertu des principes opérationnels, des critères avec

lesquels le sujet juge ses propres actions. On procède en même temps à une étude de la situation, c'est-à-dire du processus $\xi(t)$.

Le pas suivant est la construction d'un programme optimal. A la lumière de l'étude de la situation extérieure, on formule une hypothèse sur la nature de la fonction $\xi(t)$:

$$\xi(t) = \xi_0(t), \tag{5.2}$$

où $\xi_0(t)$ est une fonction du temps connue. Le problème de programme optimal peut maintenant être formulé comme un problème de commande optimale: déterminer les commandes qui permettent au système de réaliser ses objectifs avec le meilleur (le plus grand ou le plus petit selon le sens du critère) indice de qualité.

Par ailleurs, pour résoudre le problème de programme optimal on remplace souvent les équations (5.1) par une description plus simple: ceci facilite relativement les calculs. Ces simplifications sont justifiées: nous avons déjà évoqué cette question et nous y reviendrons à plusieurs reprises dans la suite.

L'étape suivante c'est la commande du programme. La trajectoire réelle du système s'écartera de la trajectoire programmée, car: 1° l'hypothèse (5.2) n'est qu'une approximation de la réalité, 2° nous avons décrit le programme par des équations simplifiées, 3° le système est attaqué par des forces qui ont été négligées, 4° nos commandes ne sont pas fidèlement réalisées. Si l'on n'intervient pas dans le caractère du mouvement, le système n'atteindra pas l'objectif de la commande. Nous devons donc envisager un système de rétroaction, c'est-à-dire un dispositif susceptible de réagir à tout écart par rapport au programme et de former de nouvelles commandes ou de modifier les anciennes de manière à minimiser l'erreur.

En résumé donc, la méthode de programmation de commande de systèmes techniques (réflectifs) se ramène à la réalisation des trois étapes suivantes.

1. Une analyse préliminaire, une étude de la situation afin de monter un scénario du processus.
2. Le calcul de la trajectoire programmée dans le cadre de ce scénario.
3. L'élaboration des mécanismes de rétroaction.

b) *Cas général de systèmes cybernétiques non réflectifs.* Considérons maintenant un système cybernétique de forme assez générale

$$\dot{x} = f(x, u_1, \ldots, u_N, \xi, t), \tag{5.3}$$

où les commandes $u_1, \ldots, u_N$ se trouvent entre les mains de sujets différents. Nous avons déjà indiqué au début de ce chapitre que l'étude de systèmes de la forme (5.3) ne pouvait être menée qu'à partir de positions subjectives. On conviendra d'effectuer l'analyse pour le sujet 1. Donc, toutes les fois qu'on dira « nous », il faut entendre le sujet 1. Dans les systèmes cybernétiques de la forme (5.3),

l'objectif n'est pas forcément un facteur exogène. Dans le même temps, le sujet qui met au point le système de commande d'un engin spatial ne choisit pas l'objectif : celui-ci lui est imposé, c'est pour lui une doctrine à l'élaboration de laquelle il ne participe pas.

REMARQUE. Cette affirmation n'est pas tout à fait exacte. En analysant les moyens de réaliser l'objectif de la commande, l'analyste peut en référer à l'échelon supérieur, par exemple en signalant l'impossibilité d'atteindre l'objectif. En d'autres termes l'analyste peut procéder à une sélection des objectifs. S'agissant des objectifs, ils se présentent comme un superobjectif, c'est-à-dire comme un objectif de niveau supérieur. Par exemple, la désignation de l'orbite d'un engin spatial est un objectif qui ne dépend pas de l'ingénieur constructeur de la fusée.

La situation est différente pour les systèmes non réflectifs, par exemple pour les systèmes économiques. En principe, les objectifs doivent être endogènes, puisque l'échelon supérieur qui les définit peut tout simplement ne pas exister, comme dans le cas d'une firme ou d'un organisme économique. Mais de quelque manière qu'il soit posé, l'objectif commence à partir d'un certain instant à jouer le rôle d'une doctrine. A vrai dire, dans les systèmes non réflectifs cette doctrine est une image subjective de l'objectif.

Nous avons déjà dit que chaque organisme poursuivait un objectif bien défini : la conservation et la consolidation de sa propre homéostasie. Or cet organisme (système) fonctionne dans des conditions qui non seulement varient constamment mais sont inconnues du sujet. Donc, il ne connaît jamais avec certitude ses propres objectifs. Par conséquent, les objectifs qu'il fixe relèvent d'une démarche subjective. En économie par exemple, l'objectif est souvent le revenu. Dans certaines conditions, par exemple dans une économie de marché, la maximisation du revenu est une garantie de stabilité. Donc, les objectifs désignés par la firme et son directeur traduisent ses vues subjectives sur les mesures indispensables à la maximisation du revenu.

Mais ces vues subjectives peuvent être imprécises (voir même inexactes). Par ailleurs la situation risque de varier avec le temps. Donc, les objectifs doivent aussi être modifiés. Par conséquent la désignation des objectifs dans les systèmes cybernétiques de type non réflectifs constitue la boucle supérieure de la rétroaction qui supervise l'homéostasie de l'organisme dans son ensemble. La désignation des objectifs dans un système économique est une sorte de mécanisme adaptatif. La répétition de la procédure de désignation des buts et les observations des résultats lèvent de nombreuses indéterminations et permettent de mieux comprendre les objectifs.

Mais en même temps varient les conditions extérieures et le caractère de l'homéostasie. C'est dans cet éternel décalage entre la conception subjective des buts et les besoins objectifs de l'organisme économique que réside la principale difficulté de la résolution du

conflit à l'échelon supérieur. L'efficacité du fonctionnement d'un mécanisme d'adaptation d'un organisme dépend grandement du caractère et de l'adéquation des procédures de désignation des objectifs. C'est ce qui distingue essentiellement les commandes de systèmes cybernétiques de celles des systèmes techniques. La formation des objectifs s'érige en problème et implique l'élaboration d'un système spécial de procédures dont nous reparlerons en traitant quelques exemples.

Cette première étape (la formation des objectifs) suppose encore la composition du scénario. Et là aussi on se heurte à une multitude de traits qui distinguent les systèmes réflectifs des non réflectifs. Dans les systèmes techniques (réflectifs) nous avons aussi à monter un scénario, mais celui-ci s'arrête à la description de la seule ambiance: la fonction vectorielle $\xi$ $(t)$. Les systèmes non réflectifs mettent en jeu d'autres sujets. Leur comportement nous est inconnu mais la possibilité de réaliser nos objectifs et ces objectifs mêmes en dépendent. C'est pourquoi il faut inclure dans le scénario nos hypothèses sur les activités des autres sujets. La composition du scénario devient un problème social complexe. Certains traits spécifiques de ce problème ont été signalés sur l'exemple de systèmes hiérarchiquement organisés et de systèmes hermeyeriens. Une fois les objectifs fixés et le scénario monté, notre système se transforme en un ordinaire système réflectif pour l'analyse duquel on peut appliquer les méthodes et principes développés pour les systèmes réflectifs.

Ainsi l'étape suivante (la deuxième) est la formation du programme. Mais les systèmes réflectifs (notamment les systèmes économiques) sont tellement compliqués que la composition du programme se fait en deux étapes.

Le calcul du programme est effectué avec un système de modèles simplifié. Ceci est aussi important que les procédures de choix des objectifs dans l'organisation de la boucle supérieure de rétroaction. Mais cette boucle ne suffit en général pas à assurer à l'organisme un fonctionnement stable. C'est pourquoi dans les systèmes économiques on distingue encore une troisième étape de planification: le calcul du plan à plus courte échéance mais avec une description plus détaillée des paramètres du système. Le programme se transforme en un plan à long terme (organisé d'une manière spéciale).

La quatrième et dernière étape est l'étape de réalisation du plan, c'est-à-dire l'élaboration des dispositifs réalisant les innombrables « petites » boucles de rétroaction.

Tel est le schéma général. Illustrons-le sur l'exemple du fonctionnement d'un organisme économique national ou régional, relevant d'un système socialiste centralisé.

c) *Formation des objectifs et prévision dans les systèmes économiques.* La méthode de programmation se présente comme une méthode universelle d'analyse des problèmes de gestion d'une économie

socialiste centralisée. C'est dire que son analyse détaillée, l'élaboration d'un appareil performant, la composition d'un de ses schémas qui permette à l'analyste de manipuler un volume d'information relativement peu élevé à chaque échelon de la hiérarchie, revêtent une signification de plus en plus importante pour la pratique.

Nous avons indiqué quatre étapes de réalisation de la méthode de programmation. Cette division est certes conventionnelle. La séparation de ces étapes n'est pas une tâche aisée. Mais quelque conventionnelle que soit cette division, elle n'en reflète pas moins de nombreux traits caractéristiques de la méthode de programmation. Essayons d'élucider le contenu de ces étapes sur quelques exemples.

Les problèmes d'étude d'une situation et de sa prévision se prêtent bien à une classification. Dans une première classe on rapporte l'étude des divers facteurs exogènes : les tendances de développement des relations internationales, le commerce extérieur, les marchés, les perspectives du progrès technique, les perspectives de développement ou d'épuisement des réserves de matières premières, les traits spécifiques de l'évolution de la biosphère, etc. Ces recherches conduisent à plusieurs scénarios. Signalons qu'il est nécessaire de procéder à une prévision qui sous un certain angle est active, c'est-à-dire que nous devons étudier non seulement le caractère de variation des facteurs exogènes mais aussi nos moyens d'action sur cette variation. La deuxième classe contient tous les problèmes liés à l'étude des possibilités de notre propre organisme économique. Les prévisions d'experts connaissent une large diffusion ces dernières décennies. C'est un important domaine d'activités qui fera l'objet d'un chapitre spécial. Mais les possibilités de l'expertise sont parfois surestimées. En réalité, plus le problème est complexe, moins sont crédibles les conclusions des experts. Et à la base de cette activité prévisionnelle on retrouve encore les modèles mathématiques. Mais leur structure doit être différente de celle des modèles utilisés pour la commande. Ces modèles doivent d'abord être relativement simples. La prévision ne définit jamais une seule trajectoire : ce n'est ni un plan ni un programme. On rappelle que l'activité prévisionnelle se rapporte à l'étape antérieure à la désignation des objectifs. Le but principal de la prévision est de faire clairement comprendre les possibilités et les perspectives de l'évolution. Nous le traiterons comme un problème de construction des ensembles permis dans l'espace des paramètres (ou critères) qui nous intéressent.

La prévision des possibilités a pour but de nous indiquer les objectifs éventuellement réalisables et ce en termes de produits finis. Mais les produits finis ne sont pas toujours les seuls à intéresser l'analyste. Les facteurs sociaux jouent aussi un rôle très important. Par exemple, toute réorganisation des proportions, toute restructuration de la production, etc., peut avoir des conséquences

fâcheuses et le sujet doit faire la part de toutes ces choses. L'analyste doit aussi penser à l'avenir. Donc, il doit étudier non seulement les possibilités limites de fabrication du produit fini, mais aussi les éventuelles capacités de production des diverses branches. D'une manière ou d'une autre, on aura affaire à toute une série d'indices (critères) $I_1, \ldots, I_k$ dans le cadre de cette étape.

Nous ne savons pas du tout formaliser les facteurs sociaux. D'où la très grande importance de l'expert qui doit passer au crible toutes les variantes et rejeter celles qui augurent des troubles sociaux.

Nous composons ensuite un modèle mathématique simplifié de notre système cybernétique, c'est-à-dire que nous décrivons le fonctionnement de l'organisme économique en termes agrégatifs. Ce système peut être ramené à la forme

$$\dot{y} = \varphi(y, u, t), \tag{5.4}$$

où $y$ est la variable de phase, $u$, la commande. Signalons que ce système fait intervenir un seul sujet et aucun facteur d'indétermination. On admet que l'étude est conduite dans le cadre d'un quelconque scénario dans lequel ont été utilisées toutes les hypothèses nécessaires. Ce modèle étudie les possibilités intrinsèques de « notre » commande.

Le problème de prévision est de construire à des dates différentes une application de l'ensemble de nos commandes $G_u$ ($u \in G_u$) dans l'espace des critères $I_i$. Ceci nous donne un ensemble $\Omega \subset I_i$ dont chaque point est associé à une stratégie.

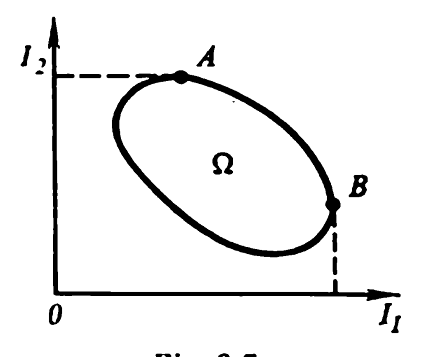


Fig. 3.7

L'ensemble $\Omega$ sera naturellement appelé *ensemble des possibilités économiques.* Nous nous pencherons plus loin sur l'estimation des programmes. Chaque programme implique une certaine dépense en biens finis, c'est-à-dire définit un point dans l'espace des critères $I_i$. Si ce point appartient à l'ensemble $\Omega$, alors les possibilités économiques de l'organisme autorisent la réalisation d'un tel programme. La portion de frontière de l'ensemble $\Omega$ qui est un ensemble de Pareto joue un rôle particulier. L'arc $AB$ de la figure 3.7 est un ensemble de Pareto. En fait, on ne s'intéressera qu'à une petite partie de l'ensemble $\Omega$ adhérant à la frontière de Pareto.

Nous avons déjà dit que les objectifs relevaient de considérations subjectives des responsables de l'organisme économique et en général étaient formulés moins en termes de produits finis qu'en termes de politique ou d'économie. La traduction de ces directives en un système de notions purement économiques et leur expression en termes de produits finis relève évidemment de la compétence des experts.

Mais ces derniers doivent appuyer leur activité sur un certain outillage, qui n'est autre précisément qu'un service de prévision utilisant une simulation grossière.

d) *Un exemple conventionnel.* Supposons qu'il est question du développement des forces productrices d'une région d'où l'on extrait une matière énergétique.

Désignons par $x_1$ les principaux fonds de la branche chargée de l'extraction de cette matière. Sa dynamique est décrite par l'équation suivante (toutes les quantités sont monétaires):

$$\dot{x}_1 = y_1 + z_1 - \varkappa_1 x_1, \tag{5.5}$$

où $y_1$ sont les investissements internes de la région, $z_1$, les investissements de l'échelon supérieur (l'Etat) ou les crédits bancaires, $\varkappa_1$, le coefficient d'amortissement.

Cette branche produit (extrait) par unité de temps une quantité de biens $P_1$ définie par la formule

$$P_1 = \alpha_1(t) f_1(x_1, L_1, Q_1), \tag{5.6}$$

où $L_1$ est le nombre d'ouvriers de cette branche, $Q_1$, la quantité de matière déjà extraite,

$$\dot{Q}_1 = P_1 \tag{5.7}$$

et $\alpha_1(t)$ est un coefficient caractérisant le rendement de la technologie (ou progrès scientifico-technique exogène). Ce coefficient montre comment à conditions égales le rendement de la production croît avec le perfectionnement de la technologie. La dépendance $\alpha_1(t)$ est définie par l'expert par expérience: c'est une extrapolation de l'expérience.

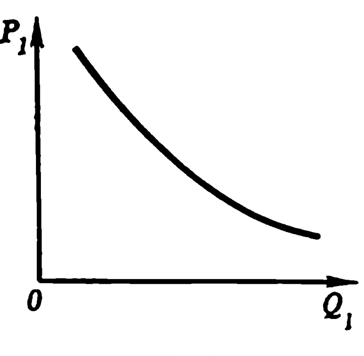


Fig. 3.8

La dépendance de $P_1$ par rapport à $Q_1$, c'est-à-dire la dépendance de la quantité de matière extraite par unité de temps par rapport à la quantité déjà extraite est une fonction convexe strictement décroissante (fig. 3.8) : la quantité de travail dépensée à l'extraction d'une unité de matière croît continûment avec l'exploitation du gisement. Ce fait est représenté sur la figure 3.8.

Dans cette région il existe encore une possibilité de production (une nouvelle technologie ou un nouveau bassin). Mais avant que cette possibilité ne commence à produire, il faut pendant un laps de temps assez long investir des capitaux qui restent improductifs durant la période de planification. Le développement de cette nouvelle branche (ou technologie) sera décrit par l'équation (en termes monétaires)

$$\dot{x}_2 = y_2 + z_2, \tag{5.8}$$

où $y_2$ et $z_2$ sont respectivement les investissements intérieurs et extérieurs.

Le reste de l'industrie de cette région sera décrit par un modèle ordinaire bisectoriel. Les fonds du premier secteur (la production des moyens de production) seront désignés par $x_3$, ceux du second (la production des biens de consommation) par $x_4$. Leur dynamique obéit aux équations

$$\dot{x}_3 = y_3 - \varkappa_3 x_3, \qquad \dot{x}_4 = y_4 - \varkappa_4 x_4, \tag{5.9}$$

et les possibilités productrices, aux fonctions de production classiques

$$P_3 = \alpha_3(t) f_3(x_3, L_3), \qquad P_4 = \alpha_4(t) f_4(x_4, L_4), \tag{5.10}$$

où $\alpha_3(t)$ et $\alpha_4(t)$ sont des coefficients caractérisant l'influence du progrès technique sur la productivité du travail, $L_3$ et $L_4$, le nombre d'ouvriers de ces deux secteurs.

Pour être en mesure de créer des fonds, la nouvelle branche a besoin d'un certain effectif. Désignons-le par $L_2$. Il est évident que $L_2$ doit être fonction des investissements $y_2 + z_2$, soit

$$L_2 = \Phi(y_2 + z_2). \tag{5.11}$$

Cette fonction est essentiellement non linéaire et est de la forme représentée sur la figure 3.9. En effet, bien avant l'exploitation des ressources allouées il faut former une équipe d'ouvriers: aucune production ne peut être réellement envisagée sans une «masse critique» d'ouvriers.

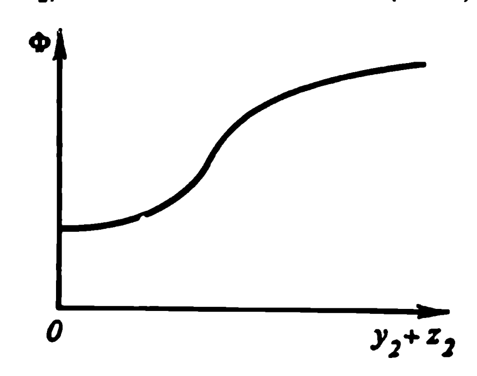


Fig. 3.9

Les investissements intérieurs $y_1$, $y_2$, $y_3$, $y_4$ de la région sont définis par le volume de la production du premier secteur:

$$P_3 = y_1 + y_2 + y_3 + y_4. \tag{5.12}$$

Enfin la main-d'œuvre est soumise à la contrainte évidente

$$\sum_{i=1}^{4} L_i(t) = L(t), \tag{5.13}$$

ù $L(t)$ est la main-d'œuvre de toute la région.

Pour que ce modèle puisse être étudié par les méthodes quantitatives, il faut le saturer d'informations.

La structure des fonctions de production est définie par des données statistiques. Les coefficients $\alpha_i(t)$ relèvent d'une prévision de l'expert. Quant à la quantité $L(t)$, elle est soit le résultat d'une

prévision d'expert, soit le résultat d'un calcul sur un modèle démographique.

Les commandes sont ici les investissements $y_i$, $z_i$ et la répartition de la main-d'œuvre $L_i$. En se donnant ces quantités en fonction du temps, on obtient un modèle fermé. Le système d'équations et de contraintes (5.5)-(5.13) nous permet de résoudre le problème de Cauchy qui consiste à définir la trajectoire de développement de la région (la quantité $x_i(t)$) satisfaisant des conditions initiales données.

Décrivons maintenant les intérêts de la région. Les responsables de la région sont intéressés en premier lieu par la satisfaction d'un certain niveau de vie des habitants de la région. Ce niveau de vie peut être défini de plusieurs façons, par exemple sous la forme intégrale

$$w_1 = \int_0^T P_4 \, dt, \tag{5.14}$$

où $T$ est l'horizon de la prévision. Au lieu de la fonctionnelle (5.14) on peut considérer la fonctionnelle

$$w_1^* = \min_t \frac{P_4(t)}{L(t)}, \tag{5.15}$$

qui caractérise la quantité de bien de consommation produite par travailleur. La condition $w_1^* \Rightarrow \max$ traduit aussi la tendance à assurer un niveau de vie élevé à la population. Par ailleurs, la production du bien $P_1$ fournit un certain revenu que la région peut consacrer à des mesures sociales, au développement de l'infrastructure régionale (construction de routes, de logements, etc.). Si toutes les quantités figurant dans le système d'équations sont exprimées en termes monétaires, le revenu tiré de l'extraction sera égal à

$$w_2 = \int_0^T (P_1 - z_1) \, dt, \tag{5.16}$$

puisque les investissements extérieurs peuvent être traités comme la somme due à la banque. Enfin, les administrateurs régionaux doivent se soucier de l'avenir. Il devient de plus en plus difficile d'exploiter les ressources avec des méthodes anciennes. Il est primordial d'implanter une nouvelle technologie ou de se lancer dans la mise en valeur d'un nouveau bassin. Donc, la direction régionale aura intérêt à utiliser des commandes telles que

$$w_3 = x_2(T) \Rightarrow \max. \tag{5.17}$$

Ainsi, le choix de la stratégie et des commandes est défini au moins par trois fonctionnelles: $w_1$, $w_2$ et $w_3$. Nous pouvons étudier

les propriétés du modèle à l'aide de problèmes de commande optimale. On pourrait proposer le scénario suivant. Posons $y_2 = z_2 = 0$. Alors $w_3 = 0$. Posons encore $y_1 = z_1 = 0$ et $L_1 = 0$. Alors $P_1 = 0$ et toutes les ressources sont dirigées vers le deuxième secteur. Cette stratégie réalise le maximum absolu de la fonctionnelle $w_1$ (l'épuisement de ses propres ressources).

L'autre cas extrême est la maximisation de la fonctionnelle $w_2$. Le problème ne sera pas non plus compliqué, puisqu'il s'agit d'un

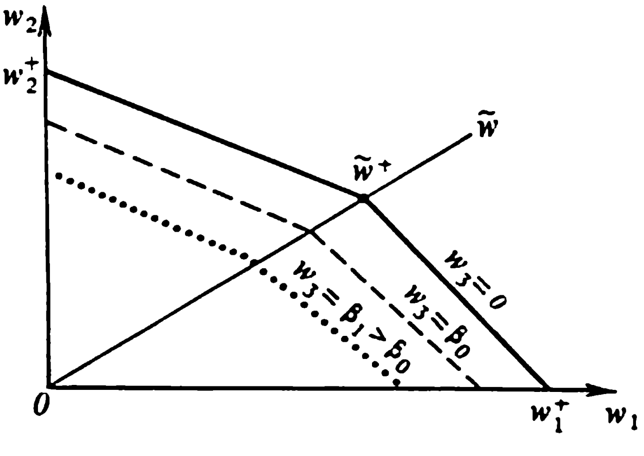



Fig. 3.10

problème à extrémité droite libre pour un système d'équations différentielles relativement facile. Dans le plan $(w_1, w_2)$, repérons les points correspondant à ces cas extrêmes: $w_1^+$ et $w_2^+$ (fig. 3.10). Il est manifeste que ces deux cas extrêmes ne sont pas réels, ils ne font qu'illustrer les possibilités limites du mécanisme économique.

Considérons maintenant une fonctionnelle de la forme

$$\tilde{w} = \gamma_1 w_1 + \gamma_2 w_2. \tag{5.18}$$

L'équation (5.18) est l'équation d'une droite dans le plan $(w_1, w_2)$ (fig. 3.10). Posons maintenant le problème

$$\tilde{w} \Rightarrow \max.$$

Soit $\tilde{w}^+$ sa solution. Relions les points $w_1^+$ et $\tilde{w}^+$ d'un côté et $\tilde{w}^+$ et $w_2^+$ de l'autre par des droites. Le quadrilatère délimité par ces droites et les axes de coordonnées décrit approximativement l'ensemble des possibilités économiques à condition que ne soient pas implantées de nouvelles branches.

Les résultats déjà obtenus sont importants: ils donnent une idée non seulement des possibilités limites de l'organisme économique, mais aussi du « coût » conventionnel de chaque fonctionnelle. Nous constatons que si une fonctionnelle varie il en va de même de l'autre. Posons maintenant

$$w_3 = \beta_0 > 0, \tag{5.19}$$

c'est-à-dire qu'on fixe le volume de production de la nouvelle branche à la fin de la période de planification. Considérons de nouveau les trois problèmes d'optimisation que nous avons étudiés pour $w_3 = 0$. Ces problèmes avec la condition (5.19) sont des problèmes isoparamétriques; quant à la condition (5.19), elle est matérialisée par la ligne en pointillé sur la figure 3.10. Posons par ailleurs

$$w_3 = \beta_1 > \beta_0$$

et portons les points de la nouvelle ligne polygonale, et ainsi de suite.

On obtient en définitive une idée assez exacte non seulement des possibilités limites de l'économie de la région, mais aussi des « sacrifices » en termes de fonctionnelles $w_1$ et $w_2$, qui assurent la création d'une réserve.

Pour obtenir des résultats plus suggestifs, on peut effectuer d'autres calculs identiques en considérant les plans $(w_1, w_3)$ et $(w_2, w_3)$.

L'exemple que nous avons traité est assez conventionnel. Mais il révèle une très importante particularité de l'étape de réalisation de la méthode de programmation que nous avons appelée étude de la situation et prévision. La prévision n'est pas un scénario. C'est une étude de l'évolution de l'organisme économique en fonction des stratégies appliquées à la réalisation de nos possibilités. La nuance est très importante, il ne s'agit pas d'une prévision globale, mais d'une prévision en fonction des commandes choisies.

REMARQUE. La construction des ensembles de possibilités économiques (l'ensemble permis) est en général un problème complexe au niveau de l'algorithme et des calculs. Mais la précision exigée n'est généralement pas élevée et la prévision réalisée est à court terme. C'est pourquoi la linéarisation du problème et la construction d'ensembles permis dans le cadre de la théorie linéaire sont assez séduisantes. Cette approche est d'autant plus justifiée qu'il existe pour les systèmes linéaires de nombreux programmes efficaces de construction et de visualisation des caractéristiques des ensembles permis. En d'autres termes, il existe un système spécial de calcul favorable à la construction des ensembles permis dans les problèmes d'assez grande dimension, de l'ordre de 10 à 15 variables (cf. [51]).

e) *Programmes*. Les programmes ont été définis comme un ensemble de mesures réalisant les objectifs de l'étape de développement considérée (étape de planification à long terme). Il est très importanl de comprendre comment se forme un programme au niveau nationar ou régional. Les programmes peuvent être de nature diverse. Par exemple:

1) Assurance d'un niveau de consommation donné.
2) Assurance d'une capacité de défense donnée.
3) Programme de développement social.
4) Grands programmes scientifiques tels que la conquête de l'espace, l'étude des fonds marins, etc.

5) Programme de mise en valeur de telle ou telle région, par exemple l'Extrême Orient soviétique ou les régions désertiques de l'Asie Centrale.

6) Programme de protection de l'environnement.

Chacun de ces programmes implique des procédures et des méthodes spécifiques d'analyse. Voyons certains d'entre eux.

*Assurance d'un niveau de consommation donné.* Il faut d'abord nous entendre sur la manière de caractériser cette notion. Il existe plusieurs procédés pour cela. Arrêtons-nous sur le plus simple.

Suivant A. Aganbéguian, le plus logique est de se donner la structure de la demande, le vecteur $d$, dont les composantes qui sont strictement positives sont normées d'une certaine façon, par exemple

$$\sum d_i = 1,$$

et le niveau de consommation $\lambda$ qui est une valeur scalaire.

Le vecteur $d$ exprimé dans des unités quelconques, par exemple en unités monétaires, indique le besoin relatif en tels ou tels biens, plus exactement en produits finis.

Les composantes du vecteur $d$ doivent faire l'objet d'une profonde analyse sociologique : il faut tenir compte de la dynamique des besoins, c'est-à-dire de l'impossibilité de modifier la structure de la demande pendant la durée du plan. De sorte qu'il est question des besoins à la fin de la période de planification. La définition du vecteur $d(T)$ est rendue difficile par le fait que le progrès scientifico-technique offre constamment de nouvelles possibilités de consommation de nouveaux produits, de nouveaux matériaux.

La détermination de la structure des besoins est un problème très ardu qui appelle une sérieuse réflexion philosophique. Faut-il ou non satisfaire tous les besoins et quels sont les plus importants ? Les réponses à ces questions ne peuvent être fournies ni par une étude de marché ni par un institut de l'opinion publique. Toutes ces mesures sont insuffisantes. Ce qu'il faut, ce sont des études complexes des problèmes de l'homme, de ses intérêts et besoins, etc.

Mais si la structure des besoins est définie, l'objectif du programme est d'élaborer une stratégie d'utilisation des ressources telle que

$$\lambda \Rightarrow \max. \tag{5.20}$$

La condition (5.20) se présente comme un critère possible d'« optimalité du plan », mais nous reviendrons sur cette question plus loin.

*Programmes de développement social, d'aide médicale, d'enseignement et autres semblables.* Le trait spécifique de ces programmes réside dans le fait que nous devons comprendre le caractère de la rétroaction : quels sont par exemple les effets d'un programme d'enseignement sur l'évolution d'une société, dans quelle mesure la réa-

lisation de ce programme assure la stabilité de cette société. La deuxième étape qui est relativement plus simple est l'estimation des dépenses en termes de produits finis.

Ces programmes peuvent être formalisés dans les mêmes termes que les programmes de consommation. L'étude des besoins en produits finis nous permet d'introduire un vecteur caractérisant la structure des besoins et l'intensité de réalisation de cette structure. Nous obtenons ainsi un système de critères du type (5.20).

*Programmes énergétiques.* Ces programmes occupent une place spéciale dans la structure de la méthode de programmation. Soulignons une très importante particularité de la méthode de programmation en économie. La planification programmée est une planification à partir des objectifs ou des produits finis, donc la production d'acier, de cuivre, de coton, bref de tout produit intermédiaire n'est pas importante en soi. C'est une conséquence du calcul des programmes. Elle n'est pas donnée mais calculée.

Mais parmi ces produits intermédiaires, il faut distinguer particulièrement l'énergie. Certes, une partie de l'énergie est utilisée par la population en qualité de produit fini. Mais la part de l'énergie électrique dépensée pour les besoins courants est dérisoire. Donc, l'énergie électrique, comme le minérai de fer ou l'acier, est un produit intermédiaire. Il n'empêche que l'énergétique occupe une place particulière et doit faire l'objet d'une étude spéciale. Et pas seulement parce que l'énergie est à la base de l'économie nationale. La création de complexes énergétiques demande plusieurs années, bien plus que tout plan à moyen terme. Donc, l'implantation d'unités énergétiques est un travail de longue haleine et le développement de l'énergétique implique des programmes spéciaux.

Supposons maintenant que les experts aient formulé une série de programmes en s'aidant de leurs services, de leurs systèmes de modèles et de leur information. Ils ont défini les mesures économiques, constructives et autres nécessaires et ont bien sûr indiqué les moyens indispensables à leur réalisation. Mais ceci ne constitue que le premier pas dans l'élaboration du programme général de développement de l'organisme économique considéré. Il est nécessaire encore de raccorder et d'harmoniser les divers programmes. En effet, les ressources nécessaires à la réalisation des programmes se trouvent toutes dans « le même sac », d'où la nécessité de les répartir entre les divers programmes.

L'harmonisation des programmes implique non seulement des procédures spéciales, mais aussi une classe spéciale de modèles fondamentalement distincts de ceux utilisés pour la prévision. Pour résoudre le problème d'harmonisation des programmes il faut au moins deux types de modèles distincts. *Primo*, chaque programme est un ensemble de travaux impliquant des ressources, des capitaux, de la main-d'œuvre. *Secundo*, nombre de ces travaux ne peuvent

être réalisés que dans un certain ordre. Voulà pourquoi le langage de la théorie des graphes est le plus propice à leur description. On dit qu'un programme est décrit si non seulement sont énumérés les travaux à accomplir, mais aussi est construit un graphe d'interdépendance de ces travaux et sont définies les ressources nécessaires à leur exécution.

Mais un programme exige des ressources. Donc, une fois qu'on a opté pour une variante de programme, il faut encore s'assurer de sa réalisabilité, de la disponibilité des ressources indispensables à sa réalisation, de l'établissement des délais d'exécution des travaux, etc. L'analyse réalisée dans le cadre de la prévision ne suffit pas à cet effet. Les calculs accomplis en étudiant les éventuelles perspectives sont des estimations préliminaires. Ce sont des jalons destinés à exclure les variantes qui sont irréalisables *a priori*. Pour fixer les délais d'exécution et estimer la réalisabilité du programme, il faut procéder à une étude bien plus poussée du processus économique. Il faut des modèles décrivant le développement et le fonctionnement des diverses branches de l'économie. Ces modèles sont bien plus détaillés que ceux utilisés pour les estimations prévisionnelles. Pour analyser la réalisabilité des programmes et estimer les éventuels délais de leur exécution, on a élaboré au Centre de calcul de l'Académie des Sciences de l'U.R.S.S. une classe de modèles, appelés $\pi$-modèles, des noms des auteurs de la première variante, A. Pétrov et Yu. Ivanilov [37].

f) *$\pi$-modèle*. Les $\pi$-modèles reposent sur la linéarité des relations d'équilibre des ressources

$$z = x - Ax - w, \tag{5.21}$$

où $x$ est le vecteur production globale de l'organisme économique (quantité de bien produite par unité de temps), $z$, la partie de production totale investie, $w$, la consommation, $A$, la matrice des dépenses directes dont l'élément $a_{ij}$ représente la quantité de bien produite par le secteur $i$ fournie au secteur $j$ pour sa production unitaire.

L'équation (5.21) s'appelle, comme nous avons déjà mentionné, équation de Léontieff. Sous la forme scalaire, elle s'écrit

$$z_i = x_i - \sum_{j=1}^{m} a_{ij}x_j - w_i, \quad i = 1, \ldots, n. \tag{5.22}$$

Au lieu de la notion de fonds utilisée dans les modèles décrits dans d), les $\pi$-modèles se servent de la notion de *puissance* $\xi$ $(t)$. On appellera puissance du secteur $i$, la quantité maximale de produits fabriqués par unité de temps. Si l'on connaît la technologie de la fabrication et le volume des fonds du secteur $i$, on peut calculer sa puissance et réciproquement.

Désignons par $\zeta_i(t)$ l'accroissement de la puissance du secteur $i$ pendant l'année $t$. La puissance du secteur $i$ au début de l'année $t+1$ sera alors

$$\xi_i(t+1) = \xi_i(t) + \zeta_i(t). \tag{5.23}$$

L'équation (5.23) décrit la variation de la puissance en fonction du temps.

Les $\pi$-modèles font encore usage des notions de puissance effective $\overline{\xi_i(t)}$ et d'accroissement de la puissance effective $\overline{\zeta_i(t)}$. Toute une série de variantes de $\pi$-modèles peut être proposée en fonction de la forme de $\overline{\xi_i(t)}$ et de $\overline{\zeta_i(t)}$. Voyons la plus simple d'entre elles.

Désignons par $\theta_i(t)$ la puissance du secteur $i$ implanté pendant l'année $t$. Ce secteur ne commence pas à produire immédiatement, il se passe un certain temps avant qu'il n'entre en service. Cette période que l'on désignera par $\eta_i$ est composée du temps nécessaire à la construction des édifices et autres aires productives, du temps pris par l'installation, le montage et le réglage des machines et enfin du temps demandé par la maîtrise de la production. Supposons que $\eta_i$ qui est une caractéristique du secteur est connue. L'accroissement de la puissance pendant l'année $t$ sera

$$\zeta_i(t) = \theta_i(t-\eta_i). \tag{5.24}$$

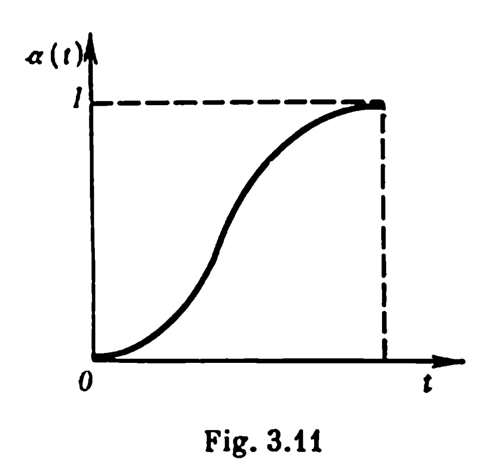


Fig. 3.11

Mais l'entreprise ou le complexe industriel commence à tourner bien avant sa mise en service totale. La mise en service des diverses unités se fait progressivement à partir d'une certaine date après le début de leur construction. Introduisons une fonction $\alpha(t)$ indiquant la part de puissance qui commence à produire au bout de $t$ années après le début de la construction. Cette fonction est schématisée sur la figure 3.11. Donc, la production maximale du secteur $i$ pendant l'année $t$ sera

$$\overline{\xi_i(t)} = \xi_i(t) + \overline{\zeta_i(t)}, \tag{5.25}$$

où l'accroissement de la puissance effective $\overline{\zeta_i(t)}$ est exprimé par la formule

$$\overline{\zeta_i(t)} = \sum_{s=0}^{\eta_i-1} \alpha_i(s)\,\theta_i(t-s). \tag{5.26}$$

Donc, $\overline{\zeta_i(t)}$ représente la quantité du bien fabriquée dans le secteur $i$

par le travail d'une entreprise dont la construction est encore en cours.

La puissance d'un secteur ne définit que ses capacités maximales. Cela ne veut nullement dire que le volume du bien fabriqué pendant l'année $t$ sera égal à la puissance du secteur. En réalité nous n'avons que l'inégalité

$$x_i(t) \leqslant \xi_i(t) + \overline{\zeta_i(t)}, \qquad i = 1, \ldots, n. \tag{5.27}$$

La quantité de bien produit est une commande et sans plus. La condition (5.27) n'est qu'une des contraintes. Nous n'avons aucune raison d'exiger que les secteurs tournent au maximum de leurs possibilités.

REMARQUE. La condition de plein rendement des secteurs est susceptible de contredire les conditions d'équilibre, puisqu'elle implique une quantité bien définie de ressources. Or ces ressources peuvent être destinées à d'autres objectifs. Dans une économie d'équilibre, on ne construira bien sûr que des unités qui sont nécessaires et qui sont pleinement chargées. Donc, « normalement » la sous-charge ne doit pas être excessive. Mais vu que nous vivons dans un monde qui est en constante mutation, vu que nous pouvons concrétiser nos objectifs avec relativement peu de temps à l'avance, vu que, enfin, les entreprises construites peuvent produire sur un intervale de temps plus long que prévu, on sera pratiquement toujours confronté à la nécessité d'une sous-charge des entreprises.

Le vecteur $z(t)$ de l'équation (5.22) est utilisé à la fabrication d'autres biens et la constitution de stocks

$$z(t) = \hat{z}(t) + q(t+1) - q(t). \tag{5.28}$$

Ici $\hat{z}(t)$ est la quantité de bien investie, $q(t)$, le volume des stocks au début de l'année $t$.

On introduit ensuite une contrainte sur la main-d'œuvre

$$\sum_{i=1}^{n} c_i(x_i) \leqslant L(t), \tag{5.29}$$

où $L(t)$ est la population active, $c_i$, la quantité de travail nécessitée par la production d'une unité de bien.

Le vecteur consommation $w$ de l'équation (5.22) est aussi un facteur de commande qui doit être minoré:

$$w(t) \geqslant w^-(t). \tag{5.30}$$

La détermination de $w^-(t)$ est un problème qui est loin d'être trivial et qui implique de profondes recherches sociologiques.

Pour que le modèle puisse fonctionner, il faut encore décrire les liens qui relient les coûts des nouvelles entreprises et des entreprises en construction.

Pour décrire les coûts des nouvelles entreprises on a adopté dans les $\pi$-modèles une hypothèse simplificatrice qu'on a convenu d'appeler *hypothèse des ratios*. Elle se formule comme suit.

Le vecteur coût $\hat{z}$ des nouvelles entreprises est défini de façon unique par le volume de production des entreprises en construction:

$$\hat{z}_i(t) = \sum_{j=1}^{n} \sum_{s=0}^{\eta_j - 1} k_{js}^{i} \theta_j (t-s), \tag{5.31}$$

où $k_{js}^{i}$, les ratios, sont des nombres fixes.

La signification de cette hypothèse est évidente. Si la construction d'une usine est en cours, elle doit se poursuivre et de plus le coût de la construction doit être strictement réglementé.

Certes, en réalité il est difficile de réglementer aussi strictement les ressources allouées à cet effet. Mais il faut convenir avec les auteurs du modèle que les plans directeurs doivent contenir des méthodes optimales de réalisation du projet et des délais optimaux d'exécution des travaux.

REMARQUE. Les $\pi$-modèles décrivent des processus qui ne sont pas markoviens. Le modèle contient un retard (cf. (5.31)), c'est pourquoi l'état initial ne définit pas de façon unique la trajectoire, d'où la nécessité de tenir compte encore du passé du processus.

Décrivons maintenant le fonctionnement du modèle. Supposons que l'état de l'organisme économique à la fin de l'année $t - 1$ est connu de même que sa préhistoire pendant les $n^*$ dernières années, où

$$n^* = \max_i \eta_i.$$

Les quantités $\zeta_i(t - 1)$ et par suite les puissances $\xi_i(t)$ seront automatiquement définies à la fin de l'année $t$. En effet, en vertu de la formule (5.24) la quantité $\zeta_i(t - 1)$ est définie de façon unique par les entreprises dont la construction a commencé pendant l'année $t - \eta_i$; quant à la quantité $\xi_i(t)$, elle est alors donnée par la formule (5.23).

Pour déterminer la puissance effective $\overline{\xi_i(t)}$, nous devons encore nous donner les quantités $\theta_i(t - s)$, c'est-à-dire les volumes de production des entreprises mises en chantier avant l'année $t$.

Donc, $\theta_i(t)$, les volumes de production des entreprises dont la construction sera entamée pendant l'année $t$, sont des paramètres libres que nous pouvons utiliser pour le développement dirigé de l'économie. Si les quantités $\theta_i(t)$ sont données, les coûts $\hat{z}_i(t)$ sont définis aussi de façon unique par la formule (5.31). Donc, la distribution des coûts entre les diverses industries est définie par la mise en chantier de nouvelles entreprises. Mais dès lors que les $\theta_i(t)$ sont fixées, leur développement et, par conséquent, la partie des

coûts destinée à la réalisation de la construction commencée, sont définis de façon unique. La production globale $x(t)$ est aussi une fonction libre. Son volume n'obéit qu'aux relations d'équilibre de Léontieff (5.22) et à la contrainte (5.27).

Considérons maintenant l'équation (5.28) qui régit les variations du volume des stocks. Si $\theta(t)$ est fixe, alors l'équation (5.28) contient encore une fonction arbitraire $z(t)$ (puisque $\hat{z}(t)$ est connue) qui peut être traitée comme un facteur de gestion des réserves.

Parmi les commandes nous devons encore classer les volumes de consommation $w(t)$ soumis aux conditions (5.30).

Donc, le $\pi$-modèle n'est pas fermé. Il contient toute une série de commandes. Dans la variante qui vient d'être décrite, ces commandes sont les fonctions suivantes:

1) le volume de la production totale $x(t)$;
2) le volume de consommation $w(t)$;
3) le volume de production des entreprises mises en chantier pendant l'année $t$, c'est-à-dire la fonction $\theta(t)$;
4) la partie de produit libre utilisée pour la création des réserves.

En agissant sur ces quantités, on peut diriger le développement de l'économie et résoudre divers problèmes d'optimisation.

Le plus important de ces problèmes est la détermination du calendrier de réalisation du programme. Certes, ainsi posé ce problème est fantaisiste en raison de sa complexité. En effet, le programme est décrit par un graphe et le problème de sa plus rapide réalisation se ramène à la composition d'un calendrier optimal des travaux dans lequel les ressources sont déterminées à partir des équations du $\pi$-modèle. Donc, la solution d'un tel problème doit être cherchée par la simulation et le calcul d'essai. Le problème de l'estimation de la réalisabilité du programme et les possibilités de combiner la résolution du problème du calendrier optimal avec celle des équations du $\pi$-modèle constituent l'un des plus importants chapitres de la théorie de la méthode de programmation en économie. Nous reviendrons à plusieurs reprises sur cette question dans le cadre de cet ouvrage.

Il existe une foule d'autres problèmes particuliers d'optimisation qui sont très utiles au stade de l'élaboration des programmes. L'un de ces problèmes est la reconstruction des proportions.

Supposons qu'en l'année $T$ nous voulions avoir une certaine structure des proportions (des puissances):

$$\xi_1(T) : \xi_2(T) : \ldots : \xi_n(T) = \lambda_1 : \lambda_2 : \ldots : \lambda_n,$$

ou

$$\xi_i(t) = \mu_i \xi_1(t), \quad \mu_i = \frac{\lambda_1}{\lambda_i}, \quad i = 1, \ldots, n. \qquad (5.32)$$

Nous pouvons alors envisager le problème de la satisfaction des conditions (5.32). Mais on peut faire un usage divers des commandes,

puisque les conditions (5.32) en définissent plusieurs. Nous pouvons en particulier exiger que

$$\xi_1 (T) \Rightarrow \max. \tag{5.33}$$

Le problème (5.33) consiste à réaliser les proportions données en économie dans les conditions d'un niveau de production maximal.

Il y a lieu de poser un autre problème d'optimisation, de rapidité celui-là: réaliser des puissances données $\xi_i (T)$ en un minimum de temps. Le $\pi$-modèle permet de résoudre encore toute une série de problèmes d'optimisation, y compris les problèmes traditionnels dans lesquels la fonction objective caractérise la consommation.

Insistons encore une fois sur l'importance des problèmes d'optimisation de la structure de l'économie. Supposons que l'organisme économique se développe en respectant ses proportions. On peut alors parler du rythme de croissance et poser par exemple le problème suivant: utiliser les commandes de sorte à maximiser le rythme de croissance. La solution de tels problèmes pourrait trouver de très importantes applications. On pourrait par exemple l'utiliser pour estimer l'effet d'une automatisation totale de la gestion. Le rythme de croissance existant est donné par le service des statistiques et est bien connu. Désignons-le par $\beta$. Soit $\beta^+$ le plus grand rythme de croissance possible. La différence $\beta^+ - \beta$ est la plus grande que nous puissions obtenir avec la structure économique envisagée. Pour augmenter le rythme de croissance, il faut modifier la structure des proportions.

Le nombre de problèmes qu'il vaut la peine de résoudre au niveau de l'élaboration des programmes peut être très grand. Chacun d'eux est porteur d'une certaine information sur les propriétés de l'organisme économique et donne plus de relief aux éventuels résultats décisionnels et aux propriétés générales du mécanisme économique.

Le succès de telles recherches dépend pour beaucoup des programmes d'exploitation du $\pi$-modèle proposé. Le Centre de calcul de l'Académie des Sciences d'U.R.S.S. a mis au point actuellement un système d'exploitation, destiné à utiliser un $\pi$-modèle de dimension peu élevée.

g) *Problème de planification et hiérarchie des modèles.* Le $\pi$-modèle décrit dans le numéro précédent était destiné à l'analyse de programmes comme instrument dans les procédures de leur formation. Ce modèle ne doit pas être de dimension très élevée (ne doit pas décrire beaucoup de secteurs), puisqu'il vise à résoudre un grand nombre de problèmes variationnels impliquant, comme on le sait, des calculs laborieux. Le nombre de secteurs peut être de 15 à 20 au plus. Mais une telle désagrégation ne suffit guère pour une planification à court terme (jusqu'à 5 ans).

Les procédures de formation des programmes déboucheront sur l'élaboration de systèmes d'indices chiffrés pour contrôler les plus

importants produits finis. Le problème majeur de la planification réelle est d'élaborer pour le niveau supérieur suivant de la hiérarchie (secteurs, groupements) des objectifs qui assureraient la réalisation des indices de contrôle. On voit ainsi apparaître des modèles plus détaillés: la panoplie des modèles devient considérablement plus riche. D'où la crainte légitime: le plan ne risque-t-il pas d'être en contradiction avec le programme?

Certes, le plan n'est pas identique au programme. Il en est le développement, la concrétisation. Le processus de planification peut révéler de nouvelles réserves et, inversement, peut mettre en évidence des « goulots d'étranglement ». C'est pourquoi les délais fixés par le plan différeront inévitablement de ceux obtenus par une analyse du programme. En fin de compte, le programme est une esquisse du plan. On peut dire que les indices de contrôle sont les jalons du plan.

D'aucun estiment que la méthode de programmation est incompatible avec les principes de planification optimale. Ce préjugé s'est si solidement enraciné qu'il faut lui consacrer quelques lignes. Quand on parle de « plan optimal », de « planification optimale », on sous-entend généralement que la fonction objectif est connue, faute de quoi ces notions n'ont pas de sens. Or le *tu autem* est dans la construction de cette fonction. Si la fonction objectif est connue, le calcul du plan est un pur problème d'optimisation. Quelque difficile qu'il soit, c'est un problème de mathématiques. Quant aux buts, ils dépendent essentiellement de circonstances non formelles basées sur une vision subjective des intérêts de l'organisme économique. La méthode de programmation est justement axée sur l'organisation de procédures permettant de fixer assez adéquatement ces buts et d'ébaucher un programme.

Le programme fixe les quantités nécessaires du bien à produire pendant l'année $T$, soit $x_i(T)$. On formule ensuite les critères pour la construction d'un système de planification optimale. Pour fonction objectif, on peut par exemple prendre:

$$J = \min_i \frac{x_i(T)}{\hat{z}_i(T)} \tag{5.34}$$

Le problème de planification optimale consiste maintenant à choisir des commandes maximisant $J$.

REMARQUE. Au premier chapitre, nous avons déjà rencontré un critère de la forme (5.34) et en avons dégagé une importante propriété: si les contraintes sont linéaires, alors le problème

$$J \Rightarrow \max$$

peut être réduit à un problème de programmation linéaire. Cette circonstance est très importante sur le plan pratique, car au stade de la planification on essaye surtout d'introduire des contraintes linéaires et de réduire le problème à un problème de programmation linéaire.

La réalisation pratique d'un schéma numérique de planification optimale donne toujours lieu à des calculs fort laborieux. La principale difficulté de ce problème est sa dimension. De nombreuses méthodes de simplification du problème de planification optimale ont été élaborées à ce jour. Prêtons notre attention à deux possibilités.

1) L'économie encaisse mal les « grands bonds »: variation brutale du volume et de la nomenclature des biens. Donc, en appliquant les méthodes itératives, on peut toujours prendre le plan quinquennal précédent en qualité de première approximation valable.

2) Il est dans la logique des choses de faire une large part aux mécanismes économiques de gestion. Ceci conduira inévitablement à une hiérarchisation de la planification qui à son tour simplifiera énormément les procédures de calcul du plan à l'échelon supérieur.

REMARQUE. Le problème du perfectionnement de la planification est d'une grande importance pratique. Il importe toutefois de bien préciser ce qu'on entend par « perfectionnement ». Il faut savoir comparer des méthodes de planification différentes. Le principal critère en la matière est la conformité du plan et du résultat final de la planification, c'est-à-dire la conformité du plan et des résultats des activités industrielles des entreprises. La non-réalisation des nouveaux indices dénote avant tout une mauvaise appréhension des spécificités des entreprises par des organes de planification.

h) *Réalisation des programmes*. La méthode de programmation, comme déjà signalé, est un système cohérent de conceptions de la gestion de systèmes cybernétiques complexes oligopoles, y compris de systèmes économiques. L'analyse préliminaire (la prévision), la confection des programmes, la planification — toutes ces étapes doivent s'achever par la réalisation de programmes et plans, c'est-à-dire par la création de mécanismes spéciaux de réalisation (du genre pilote automatique). Transposée aux systèmes économiques, cette proposition signifie que les programmes doivent être projetés en même temps que les mécanismes économiques.

Ces problèmes complexes sont devenus depuis peu l'objet d'une analyse scientifique. Ils couvrent un vaste champ d'activités: étude de la structure des systèmes de gestion, choix du rapport optimal des droits et de la responsabilité des divers chaînons et échelons de la hiérarchie, élaboration des principes de conception de nouvelles et de perfectionnement des anciennes entreprises, etc.

Actuellement il est beaucoup question d'expérience économique. Mais quelque importante que soit la place réservée à l'expérience dans toute recherche, celle-ci est toujours insuffisante, surtout en économie. L'expérience économique non seulement est coûteuse, non seulement touche au destin des gens, mais elle est très difficilement transposable à d'autres objets, et ses conclusions, généralisables. Elle est toujours concrète. C'est pourquoi les recherches théoriques sont d'une nécessité absolue dans ce domaine.

Au premier chapitre, de même qu'ici, nous avons abordé certaines spécificités de la gestion dans les systèmes hiérarchiques. Il me semble que l'approche informationnelle de l'analyse du fonctionnement des systèmes économiques qui a été décrite, combinée à l'examen de la situation concrète et à des investigations sociologiques peut servir de base à la promotion de cette importante branche des mathématiques de l'économie qui sera consacrée aux problèmes de création de mécanismes économiques.

Certains auteurs estiment que la conception de tels mécanismes est caractéristique d'une économie planifiée, puisque dans l'économie de marché, l'archétype des mécanismes — le marché — est spontané. Cette affirmation n'est pas tout à fait exacte. Certes le marché est un phénomène spontané sur les lois duquel l'homme a difficilement prise. Une prise difficile mais pas impossible. La monopolisation de la production industrielle et du commerce a conduit à l'apparition de toute une série de situations nouvelles. En particulier, divers mécanismes ont commencé à se former de façon dirigée, et j'insiste bien sur le mot « dirigé ». C'est pourquoi les conceptions des mécanismes économiques développées dans ce chapitre et les principes de leur analyse peuvent être appliqués à l'étude d'une économie capitaliste.

Par ailleurs, les mécanismes de l'économie planifiée n'ont pas tous et de loin une vocation bien déterminée. Nombre de mécanismes qui ont joué un grand rôle dans la définition des processus économiques internes ont vu le jour spontanément. Il y a plusieurs causes à cela. La plus importante est la complexité de la réalité objective à laquelle est confronté l'analyste. Le fonctionnement de tout système économique est régi par des lois objectives et par des contradictions qui lui sont propres. La connaissance imparfaite de ces lois et contradictions se solde par un fonctionnement différent de celui attendu à cause précisément des effets imprévus non programmés.

La discussion de telles questions est très intéressante et très importante pour la pratique. Mais dans l'ensemble ce sujet outrepasse le cadre de notre ouvrage. On retiendra essentiellement la conclusion suivante : les projets des mécanismes réalisant un objectif ou un programme sont tout aussi importants que le programme en question ou la planification. Les mécanismes parfaits de gestion économique sont aussi vitaux à un organisme que le pilote automatique l'est pour un engin spatial. Un appareil admirablement construit, doté de tous les moyens nécessaires, avec un objectif correctement posé et un programme optimal bien élaboré n'atteindra jamais le but assigné s'il est équipé d'un mauvais système de correction.

L'unité de l'objectif, du programme (et du plan) et des mécanismes de réalisation constitue l'un des principes sacro-saints des systèmes, principe qui doit trouver place dans une théorie élaborée. Des pas dans le sens de la création de cette théorie ont été accomplis au cours des 20 dernières années et ont été reflétés dans cet ouvrage.

i) *Sur l'historique de la méthode de programmation.* Une abondante littérature a été consacrée à la méthode de programmation depuis son apparition en raison de son retentissement. Ces travaux portent des jugements différents quant à son origine et sa paternité.

La plupart d'entre eux sont à mon sens le fruit d'un malentendu. Le fait est que l'apparition de la méthode de programmation est souvent liée à la procédure de programmation et de budgétisation qui du temps de J. F. Kennedy fut mise à la disposition du budget militaire américain par le secrétaire d'Etat à la Défense d'alors McNamara. La méthode d'établissement du budget introduite au ministère de la défense des U.S.A. dans les années 60 répondait à un problème assez spécifique: l'ordonnancement des dépenses des divers corps d'armées assumant des missions générales. La méthode de McNamara se fixait donc pour objectif la création de procédures liant les buts nationaux à la structure du budget militaire. Donc, de même que dans la méthode de programmation décrite plus haut, toute la planification du budget était conduite à partir des objectifs finaux. L'apparition du terme « méthode de programmation » aux U.S.A. a de toute évidence retenu notre attention et a contribué à l'implantation de cette méthode chez nous. Mais les idées et procédés de résolution des problèmes, proposés aux U.S.A. du temps de McNamara, n'ont pratiquement eu aucune incidence sur le développement de cette méthode et sur ses aspects gnoséologique et instrumental.

Je me suis appliqué à montrer que la méthode de programmation est un système de points de vue, une conception unique et cohérente de gestion de systèmes cybernétiques de forme très générale. C'est un système qui permet d'unifier les méthodes formelles et non formelles d'analyse pour la gestion des systèmes non réflectifs dans des situations complexes, lorsque l'environnement et les conditions de l'homéostasie des organismes varient constamment et imposent *ipso facto* une restructuration des objectifs.

Les bases formelles de la méthode de programmation ont été jetées en théorie des systèmes de type réflectif et tout d'abord dans les systèmes techniques fonctionnant dans des conditions non stationnaires. Les notions de « programme », « programme optimal », « optimisation en deux étapes » et autres, importantes pour l'appréhension de l'essence de la méthode de programmation ont été forgées et utilisées en mécanique des vols spatiaux à l'aube de l'ère des ordinateurs. Dès la fin des années cinquante toutes les conditions étaient réunies pour transposer ce système de conceptions du cadre technique de la théorie de la commande au cadre économique de la théorie de la gestion et de l'adapter aux nouveaux problèmes. Il fallait un prétexte qui obligeât les spécialistes en théorie de la commande à se tourner vers ces problèmes. Ce prétexte leur fut fourni par l'échec de l'application de l'idée de la planification optimale en U.R.S.S.

Les prémisses et principes de la planification optimale et les systèmes préconisés par divers auteurs ont fait l'objet de débats très animés au début des années soixante. Les principes d'optimalité étaient le point de mire de la critique. A nous, spécialistes de la théorie de la commande et de la recherche opérationnelle, il nous semblait que la difficulté majeure consistait justement à construire la fonction objectif. Ceci devait précisément constituer l'épine dorsale de l'analyse économique. Quant aux méthodes mathématiques de résolution des divers problèmes, elles furent reléguées au second plan. On assista à une confrontation des idées des « traditionnalistes »

de la recherche économétrique et des « pragmatistes » de la théorie de la commande et de la recherche opérationnelle.

La seule analogie avec la théorie de la commande ne suffisait pas pour élaborer un autre système de conceptions. Il fallait donner à ce système un contenu économique. A cet effet, il était nécessaire tout d'abord de puiser dans la réalité, dans l'histoire concrète du développement de l'économie soviétique.

Et bien évidemment, l'histoire nous fournit des exemples suggestifs à profusion.

L'exemple le plus marquant, celui où les principaux contours de la méthode de programmation ressortent avec le plus de relief, correspond à l'étape initiale de restauration de l'économie soviétique après la guerre civile.

Le Parti décréta une restauration de l'industrie qui servît de point de départ à un rapide essor du pays, à sa transformation de pays agraire en pays industriel.

C'était une doctrine, une directive du Parti. Comment la réaliser ? Toute une série de mesures, ou de programmes, en employant la terminologie actuelle, furent proposées.

Le plus brillant fut manifestement le plan GOELRO *) qui fut qualifié par Lénine de deuxième programme du parti.

Ainsi le plan GOELRO n'était qu'un programme parmi d'autres. Il spécifiait certaines mesures nécessaires à la restauration énergétique. Et pas seulement de restauration. En effet, avant la révolution la Russie disposait d'un grand nombre de petites centrales électriques essentiellement thermiques. Il fut décidé de les remplacer par un réseau de grosses centrales régionales qui à la longue deviendraient la base d'un système énergétique national unique.

Ce qui frappe à l'étude des matériaux liés à l'élaboration du programme GOELRO, c'est la minutie et la finition des détails. Il était clair que la tâche principale de la planification et de la gestion consistait non pas à optimiser une fonction objectif utopique, mais bien à mettre sur pied des procédures transformant les directives du parti (les doctrines) formulées en haut lieu en mesures économiques concrètes. Et ce système de procédures devait couvrir toutes les étapes — de la création des outils scientifiques exigés par l'élaboration de la doctrine à sa réalisation.

En conséquence, le terme « méthode de programmation » est né en U.R.S.S. par suite de l'élaboration de procédures de réalisation des programmes du parti.

Les techniciens participant à la confection de ce système de points de vue qui allaient à l'encontre de la théorie de planification optimale, théorie qui tenait le haut du pavé à l'époque, étaient loin de connaître les idées de McNamara et en tout état de cause ne leur accordèrent aucune importance.

---

*) Plan d'Etat d'électrification de la Russie (*Note du traducteur*)

## § 6. Simulation et expérience sur ordinateur

Nous avons à maintes reprises évoqué les termes de « simulation », d'« utilisation d'un modèle en régime de simulation », etc. Ces notions sont intuitivement évidentes. Mais dans ces notions apparaît en filigrane une chose plus importante que l'intuition. Aujourd'hui la simulation est l'une des méthodes les plus importantes et les plus efficaces dont dispose l'analyse des systèmes. Ces questions feront l'objet de ce paragraphe.

a) *Méthode de Monte-Carlo.* Le terme « simulation » a fait son apparition dans la littérature anglaise au début des années soixante dans le cadre de l'application de la méthode de Monte-Carlo à l'étude de processus dépendant de fonctions ou de paramètres aléatoires. Considérons l'équation

$$\dot{x} = f(x, t, \xi), \tag{6.1}$$

où $x$ est la variable de phase, $\xi$, un paramètre aléatoire dont la loi de probabilité est connue. Soit le problème de Cauchy pour cette équation: on demande la trajectoire du système (6.1) qui vérifie la condition

$$x(0) = \eta, \tag{6.2}$$

où $\eta$ est une variable aléatoire de loi de probabilité connue. La trajectoire de phase du système (6.1) sera dans ces conditions une fonction aléatoire du temps et $x(T)$, une quantité aléatoire. Supposons qu'on s'intéresse à la répartition de cette quantité.

REMARQUE. Il existe une foule de problèmes techniques qui se réduisent directement à un problème de cette nature. Citons notamment les problèmes de balistique: sachant les paramètres de la répartition des erreurs aléatoires de réalisation des conditions initiales ($\eta$) et les paramètres des perturbations extérieures ($\xi$), calculer les paramètres de l'ellipse de dispersion.

Si la fonction $f$ est linéaire, ce problème ne soulève aucune difficulté particulière. Si $f$ est essentiellement non linéaire, nous ne disposons d'aucune méthode universelle de résolution. Le problème de déterminer les paramètres de la répartition du vecteur aléatoire $x(T)$ sachant les répartitions des variables $\xi$ et $\eta$ fait partie de la classe des problèmes dits de filtration non linéaire. Pour l'instant on n'a pas encore réussi à élaborer des méthodes tant soit peu satisfaisantes pour résoudre cette classe de problèmes.

Mais le calculateur a toujours la possibilité de se rabattre sur la méthode suivante dite de Monte-Carlo.

Soit donnée une fonction

$$y = \varphi(\xi), \tag{6.3}$$

où $\xi$ est une variable aléatoire de loi de probabilité connue. Un générateur de nombres aléatoires nous permet de construire une suite de

nombres aléatoires

$$\xi_1, \xi_2, \ldots, \xi_N \tag{6.4}$$

possédant la loi de probabilité désirée. La formule (6.3) nous donne ensuite la suite de valeurs

$$y_1 = \varphi(\xi_1), \qquad y_2 = \varphi(\xi_2), \ldots, \tag{6.5}$$

qui sera aussi une suite aléatoire. Si l'on effectue un assez grand nombre de calculs, on peut par un traitement de la suite (6.5) déterminer les paramètres statistiques de la variable aléatoire $y$ avec n'importe quelle précision et trouver ensuite la loi de probabilité annoncée.

Il est évident que la méthode de Monte-Carlo peut être appliquée à la résolution du problème formulé au début du paragraphe.

On peut toujours à l'aide du générateur de nombres aléatoires définir une suite de nombres

$$\begin{matrix} \xi_1, \xi_2, \ldots \xi_N, \ldots, \\ \eta_1, \eta_2, \ldots \eta_N, \ldots, \end{matrix} \tag{6.6}$$

résoudre pour chaque couple $(\xi_i, \eta_i)$ le problème de Cauchy (6.1), (6.2) par le procédé classique et déterminer la suite

$$x_1(T), x_2(T), \ldots$$

Cette méthode d'analyse se généralise aisément aux cas plus compliqués où l'équation (6.1) contient non seulement des paramètres aléatoires, mais aussi des fonctions aléatoires.

L'extension de la méthode de Monte-Carlo aux systèmes dynamiques a reçu le nom de « simulation ».

Signalons deux circonstances. Tout d'abord, la méthode d'analyse décrite peut être considérée comme un procédé de détermination empirique sur machine des paramètres du processus aléatoire $x(t)$.

Ensuite, cette méthode peut être traitée comme une analyse du processus dynamique (6.1) à l'aide des calculs d'essai.

Ces deux circonstances ont joué un rôle non négligeable dans le développement des techniques de simulation.

b) *Sur la notion de système de simulation.* La méthode d'analyse des processus complexes à l'aide des calculs d'essai, qui a été décrite plus haut, s'est largement répandue dès le milieu des années soixante aussi bien en U.R.S.S. qu'à l'étranger.

On peut indiquer deux causes du succès des idées de la simulation.

La première est celle qui a conduit à l'apparition de l'analyse des systèmes, savoir la grande complexité des systèmes que l'homme a eu à manipuler ces dernières décennies dans ses activités techniques, économiques, militaires.

J'ai déjà essayé d'attirer l'attention du lecteur sur la distinction

entre les problèmes de l'analyse des systèmes et ceux de la théorie de la recherche opérationnelle et de la théorie de la commande. Dans ces derniers l'objectif était traditionnellement donné et il fallait chercher les moyens de le réaliser. Au contraire dans les nouveaux problèmes qui se posaient constamment aux chercheurs opérationnels, l'objet de la recherche était les objectifs en question.

Il est évident que ce problème est loin d'être un problème de mathématiques pures et ne peut donc être résolu sans l'intervention d'experts. Mais les experts ne pourront pas à eux seuls surmonter cette avalanche d'information à analyser. Le tandem mathématicien-expert est donc une exigence du temps. Nous l'avons déjà dit et répété. Et la première chose indispensable à la réalisation de telles idées est une bonne organisation des calculs d'essai : l'expert doit clairement appréhender le caractère du processus étudié, le degré de sa « commandabilité », ses possibilités limites (ensembles permis) c'est-à-dire mettre sur pied un plan d'expériences. A cet effet, il faut construire des modèles simulant le processus étudié. A l'aide de ces modèles et des calculs d'essai l'expert recueille les informations qui l'aideront à choisir sa stratégie. Les spécialistes ont très vite saisi tout le bénéfice qu'ils pouvaient tirer des ordinateurs.

Donc, la base de la simulation (par simulation on comprendra l'analyse par des calculs d'essai) est le modèle mathématique. Si le modèle est défectueux ou inexact, il est aberrant de parler de simulation. Le célèbre naturaliste du siècle dernier Th. Huxley a dans une remarquable boutade comparé les mathématiques à une meule qui écrase n'importe quel produit. Mais on ne fait pas de la farine à partir de l'ivraie même avec les plus belles meules. On peut en dire autant du modèle : celui-ci doit être de bon aloi et refléter correctement la réalité, sinon son étude se transforme en un simple divertissement mathématique.

Mais cela ne suffit pas. L'information de départ doit être sûre et pertinente. C'est l'évidence même, mais ce n'est pas encore tout. Il faut un certain service. La panoplie de modèles doit être facilement accessible à l'analyste, les diverses variantes doivent passer assez facilement, le système de visualisation (graphique à l'aide d'un traceur de courbes ou d'un display, ou numérique) doit bien fonctionner. Tout doit être simplifié à l'extrême, de l'introduction de toute nouvelle information au passage à une nouvelle variante, etc. En d'autres termes, pour être mené à bien, le processus de simulation implique la mise au point d'un système spécial : c'est ainsi qu'est né le terme *système de simulation*. Donc, le système de simulation est un ensemble de modèles imitant le processus étudié grâce à un corps spécial de programmes auxiliaires et des informations permettant d'effectuer assez rapidement les calculs d'essai. Le système de simulation obéit à un certain principe architectonique et son mode d'emploi doit exclure toute ambiguité.

Par ailleurs, les systèmes de simulation doivent leur création à la nécessité de résoudre des problèmes complexes d'optimisation (en particulier, l'optimisation de constructions ou de solutions économétriques). Supposons qu'on ait à résoudre le problème suivant: trouver un élément $x$ (un vecteur ou une fonction vectorielle du temps) qui maximise une fonctionnelle $W$:

$$W(x) \Rightarrow \max_{x \in X}.$$

Supposons par ailleurs que l'on connaît son algorithme de résolution et que celui-ci est convergent et stable. Sur le plan mathématique il n'y a apparemment aucune entrave à la résolution de ce problème. Mais cet algorithme ne servira pas à grand chose si le temps nécessité par les calculs dépasse de loin les limites qui nous sont imparties.

Cette situation se présente généralement dans les problèmes d'emploi du temps. La résolution de ces problèmes est souvent un élément de commande opérationnelle, par exemple la conduite de vaisseaux dans un port ou d'autres travaux. Le dispatcher doit prendre une décision en l'espace de quelques minutes, ou à la rigueur de quelques heures. Or la résolution exacte de tels problèmes peut demander des mois, voire même des années de travail continu d'un ordinateur. Dans ces conditions, l'analyste n'a qu'un seul et unique recours: les méthodes intuitives et euristiques. Mais, avant de prendre une décision il doit s'assurer qu'elle est satisfaisante, il doit la supputer, la comparer à d'autres solutions. A cet effet, il doit calculer $W(x)$ et éventuellement d'autres critères accessoires. On se trouve donc de nouveau devant la nécessité d'effectuer dans un court délai un grand nombre de calculs d'essai.

Comme la solution euristique proposée par le dispatcher risque de ne pas être assez bonne, il faut avoir la possibilité de vérifier d'autres variantes, de les comparer, et de disposer des moyens pour les préciser. Et même si le système de simulation est un instrument fort complexe, on voit mal comment l'analyste pourrait s'en passer dans de telles situations.

Signalons que l'usage des systèmes de simulation dans les systèmes automatiques de commande est de plus en plus nécessaire au fur et à mesure que de systèmes de traitement des données ils se transforment progressivement en systèmes de prise de décision. Dans les problèmes d'automatisation des projets de systèmes techniques complexes, ils sont devenus le principal outil de vérification et de comparaison des diverses variantes et même souvent de leur création.

La simulation prise dans son acception large est en train de devenir le principal instrument de l'analyse des systèmes.

Il est encore une cause qui explique que les principes de la simulation et l'utilisation des systèmes de simulation soient en vogue et

fassent l'objet d'une analyse détaillée, c'est l'apparition des ordinateurs de la troisième génération.

Du point de vue de l'usager, l'avantage des ordinateurs de la troisième génération n'est ni dans les éléments de base (que les calculs se fassent à l'aide de transistors ou de circuits intégrés, cela lui est égal), ni dans la rapidité, ni dans la capacité de mémoire. Certains ordinateurs de la deuxième génération possédaient une mémoire et une rapidité satisfaisantes. L'avantage donc des ordinateurs de troisième génération sur leurs prédécesseurs est dans le service terminal, facilitant énormément l'entrée, la sortie et la visualisation de l'information, l'organisation du travail en temps partagé, etc.

L'implantation des ordinateurs de troisième génération a permis la réalisation technique des principes de simulation, principes qui étaient connus depuis belle lurette. Mais c'est cette nouvelle génération qui a rendu possible le tandem homme-machine et l'unification des méthodes rigoureuses d'analyse et des méthodes euristiques.

Ce processus se poursuit encore. Les ordinateurs personnels (ou professionnels) construits dans les années 80 jouent un rôle particulièrement important dans l'évolution des méthodes de dialogue dans l'étude des projets et les prises de décision à l'aide de systèmes de simulation. Ces ordinateurs mettent les systèmes de simulation à la portée de personnes n'ayant pratiquement aucune notion de mathématiques.

c) *Remarque sur l'intellect artificiel.* Ce terme est en vogue ces dernières années. Il a été forgé par les ingénieurs dans le cadre de l'utilisation des ordinateurs à la résolution de problèmes non traditionnels pour les mathématiques.

Intellect artificiel et système de simulation sont en fait synonymes. Un exemple d'architecture d'un tel système est représenté sur la figure 3.12. Ce système possède de nombreux traits qui appellent quelques commentaires.

L'utilisateur n'est pas obligatoirement un spécialiste en mathématiques. Ce peut être un ingénieur, un constructeur, un économiste, un écologiste, un administrateur et on ne peut par conséquent exiger de lui des connaissances poussées en mathématiques ou même la connaissance des langages de programmation. Plus le système sera facile à manipuler (c'est-à-dire plus il sera sophistiqué du point de vue du mathématicien), plus il aura des chances de conquérir ses clients. Les systèmes de simulation doivent précisément viser des usagers non initiés. Donc, l'élaboration du langage de l'usager est la première importante étape de la création du système. Mais comme les premiers systèmes de simulation ont été construits par les mathématiciens pour leur usage personnel, le problème du langage ne s'est pas posé pendant longtemps. Dans les travaux des ingénieurs portant sur la création d'un intellect artificiel, le langage de l'usager était au premier rang des problèmes.

Le programme de gestion (nous l'appelons souvent système opérationnel exogène pour le distinguer du système opérationnel endogène — du système opérationnel de l'ordinateur) est l'élément du système de simulation qui est le principal responsable du dialogue. La construction du système opérationnel exogène dépend particulièrement de l'ordinateur sur lequel est réalisée la simulation.

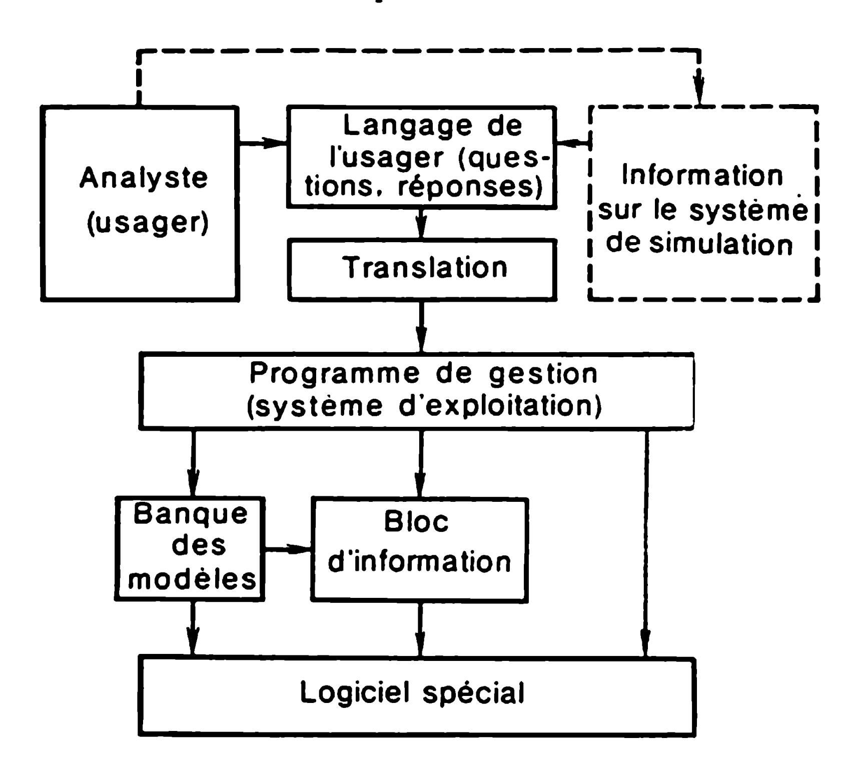



Fig. 3.12

Au programme de gestion sont reliées une banque de modèles et une banque de données (bloc informationnel). Sur la figure 3.12 on a tracé une flèche de la banque des modèles au bloc informationnel. Cette flèche symbolise la priorité des modèles sur l'information. La remarque que nous venons de faire est dictée par la raison suivante. Avant l'apparition des ordinateurs, on avait une certaine notion de l'information nécessaire dans les problèmes de gestion et d'élaboration des projets. Le passage à une nouvelle technologie d'élaboration des projets est basé surtout sur l'utilisation des modèles. C'est précisément l'analyse des modèles qui a permis de mettre en lumière le parti qu'on pourrait tirer des ordinateurs. Mais si l'on s'appuie sur l'information traditionnelle, l'apport des ordinateurs sera pratiquement nul, car cela signifiera que nous utilisons des modèles pour lesquels cette information est exhaustive, c'est-à-dire des modèles simples, semblables à ceux utilisés antérieurement. Les nouvelles techniques d'élaboration des projets impliquent des raisonnements nouveaux et une information d'un type nouveau.

Les mathématiciens qui ont commencé à étudier les problèmes de simulation partaient des processus et de leurs modèles. Le point fort de leurs activités était la création de modèles et de procédures (d'algorithmes) pour les étudier. Ce n'est que plus tard qu'ils furent confrontés à la nécessité d'un système d'exploitation, de programmes de gestion, d'un langage spécial, etc. Les ingénieurs suivaient la démarche inverse. Ce qui les intéressait surtout, c'étaient l'élaboration de systèmes d'exploitation, des programmes de gestion, des banques de données, etc. Donc, dans la simulation, toute l'attention était concentrée sur les modèles, et dans la construction d'un intellect artificiel, sur le système opérationnel, les langages, les banques de données, etc. Mais en fait cela revient au même, car ce sont là des facettes différentes d'un problème, immense et infiniment important pour l'humanité, d'unification de l'intellect humain avec un formalisme dont la réalisation est accessible aux mathématiques armées d'ordinateurs.

Sur la figure 3.12, le bloc en pointillé appelle quelques commentaires.

Les systèmes de simulation actuellement en fonctionnement ou en passe de création sont des constructions fort compliquées dont l'étude exige non seulement des connaissances très poussées mais aussi du temps. L'usager n'a pas besoin de connaître tous les détails de la construction. L'ingénieur qui imagine le moteur d'un avion n'est pas censé connaître toutes les subtilités du métier de constructeur du fuselage et inversement. Mais en revanche chacun d'eux doit appréhender nettement toutes les possibilités qui lui sont offertes par le système de simulation pour ses propres activités.

Ces fonctions sont remplies par le bloc « information sur le système de simulation ». Ce bloc est un système informationnel arborescent aux extrémités duquel sont décrits les problèmes qui intéressent le projeteur. La création d'un tel bloc est dictée par une autre raison. Le système de simulation (s'il est récursif) doit nécessairement être évolutif. Il doit non seulement permettre le remplacement de certains modèles par d'autres, des vieux algorithmes par de plus perfectionnés, mais aussi permettre d'élargir le diapason de ses fonctions, de perfectionner le langage, etc. L'information sur le système de simulation, sur tous les détails de ses performances et de sa construction est très utile à cet effet. Et de plus, l'accès à cette information doit être facile.

L'analyse et la comparaison des travaux portant sur l'intellect artificiel et sur la simulation mettent en évidence une autre différence dans les activités des ingénieurs et des mathématiciens.

Les ingénieurs spécialisés dans l'utilisation des ordinateurs essayent de créer des systèmes opérationnels universels qui conviendraient à tous les modèles et satisferaient un large éventail d usagers. Ceci explique que les systèmes s'avèrent assez compliqués

et soient difficiles à manipuler dans les problèmes pratiques. Au contraire, les mathématiciens qui s'occupent de la simulation tentent d'adapter au maximum le service tout entier à la résolution du problème de simulation étudié, qu'il s'agisse d'un système automatique de construction d'un avion ou d'un système de modèles destinés à analyser la guerre du Péloponnèse.

L'évolution de cette orientation dépendra beaucoup de la manière dont se combineront ces deux tendances contradictoires.

d) *Simulation et expérience sur ordinateur.* La simulation utilise les ordinateurs comme des instruments d'expérience. A proprement parler, cet usage date des années cinquante. A l'époque il fallait étudier des phénomènes qu'il était impossible de reconstituer sur terre. Par exemple, la pénétration d'un engin cosmique dans l'atmosphère à une vitesse de l'ordre de plusieurs kilomètres par seconde. Pour étudier de tels processus on ne pouvait procéder à des expériences ni en laboratoire, ni sur le modèle. L'analyste n'avait qu'une seule solution : mettre au point un modèle mathématique parfait et déduire les caractéristiques nécessaires du processus étudié par un plan d'expérience.

Cet usage direct des ordinateurs en tant qu'instruments d'expérimentation était la règle dans les années cinquante et au début des années soixante. Mais dès cette époque se dessinait une nouvelle orientation dans le domaine de l'expérimentation sur ordinateur : l'utilisation des ordinateurs non seulement pour étudier les propriétés du modèle, mais aussi pour fabriquer un modèle mathématique.

Au début de ce paragraphe nous avons signalé la base phénoménologique de toute nouvelle théorie. Certes, toute théorie s'appuie sur une expérience. Mais cette expérience, surtout si l'expérimentateur avance à tâtons, risque d'être très imparfaite. L'information qu'elle fournira à l'expérimentateur sera encore insuffisante pour que le modèle mathématique conçu reflète adéquatement la réalité. Supposons tout de même que cette expérience première a permis de formuler le modèle mathématique.

L'étape suivante est l'analyse numérique du modèle. Cette analyse nous permet de dégager de nombreuses propriétés. Certaines d'entre elles auront été observées par l'expérimentateur, d'autres seront nouvelles. Cependant il faut se garder des affirmations trop hâtives et encore plus des pseudodécouvertes de nouvelles propriétés. Le modèle mathématique qui a permis de mettre ces traits en évidence est loin d'être parfait et les résultats des calculs sont peu crédibles. Mais aussi imparfaites qu'elles soient, ces expériences sont très utiles, dans la mesure où elles suggèrent de nouvelles idées, et permettent à l'expérimentateur d'affiner son appareillage et certains détails du processus, etc. Cette procédure débouche sur un nouveau modèle mathématique. Une chaîne itérative se forme : un dialogue entre le mathématicien et l'expérimentateur. La procédure décrite

a donné de nombreux et intéressants résultats en physique et dans les sciences techniques.

J'ai employé le mot « dialogue » *ad rem*. Toute investigation est dialogue : des questions sont posées, il leur faut trouver des réponses. Et le physicien expérimentateur et le mathématicien qui a conçu et analysé numériquement son modèle, poursuivaient un but bien défini. Ils posaient des questions : le premier, à la nature en se basant sur ses calculs sur ordinateur, le second, à son modèle en se référant aux résultats de l'expérience. Mais ce dialogue n'était pas structuré.

Ces vingt dernières années, dans les sciences de la nature, en économie et dans les sciences techniques, on assiste à la formation d'un nouveau type d'expérience qui fait la part non seulement à l'étude du phénomène mais aussi au choix de ses caractéristiques ou à l'analyse de nombreuses relations. Ce nouveau type d'expérience implique des systèmes de simulation dotés d'un système opérationnel et d'un terminal sophistiqués.

Depuis la fin des années soixante, les techniciens du Centre de calcul de l'Académie des sciences d'U.R.S.S. ont consacré d'importants efforts au développement de ce nouveau type (cf. [5]) qui exige un dialogue perfectionné et l'élaboration d'un système de procédures spéciales.

## § 7. Modèle de fixation des pénalisations pour la pollution de l'environnement

Voyons en conclusion un exemple de système cybernétique décrivant les rapports entre l'Administration d'une région et les entreprises disséminées sur son territoire. Ces entreprises ont des activités qui polluent l'environnement (par exemple l'eau) et l'Administration régionale doit élaborer une stratégie dans ses rapports avec elles. Analysons un schéma simplifié de cette situation et montrons la place et la nécessité d'un système de simulation dans de telles recherches.

Désignons par $\Phi_i$ $(i = 1, 2, \ldots, N)$ les fonds des entreprises. Leurs variations seront régies par les équations suivantes :

$$\dot{\Phi}_i = Y_i - k_i\Phi_i, \qquad (7.1)$$

où $Y_i$ sont les investissements, $k_i$, les coefficients d'amortissement.

La production de chaque entreprise par unité de temps est

$$P_i = F_i(\Phi_i), \qquad i = 1, \ldots, N, \qquad (7.2)$$

où $F_i$ est la fonction de production.

Les entreprises produisent des biens mais aussi des déchets. Désignons par $\pi_i$ le flot des matières polluantes :

$$\pi_i = f_i(P_i, V_i), \qquad (7.3)$$

où $V_i$ sont les dépenses des entreprises pour le perfectionnement des technologies ou l'épuration de l'eau dans les conditions de l'usine.

On conviendra d'étudier une situation mettant en jeu $N+1$ agents: $N$ entreprises et le Représentant des intérêts de l'Administration régionale. Ce dernier a le droit de pénaliser les entreprises. On admettra que la pénalisation $w_i$ est proportionnelle à la quantité de déchets rejetés:

$$w_i = c\pi_i = cf_i\,(P_i,\ V_i). \tag{7.4}$$

Le coefficient de pénalisation $c$ dépend naturellement de l'Administration régionale. Les autres commandes $V_i$ et $Y_i$ sont toutes à la disposition des entreprises. L'Administration n'a pas le droit de s'ingérer dans les affaires des entreprises.

On conviendra pour simplifier que les investissements sont endogènes et que tous les capitaux sont utilisés pour couvrir les frais de développement, de construction d'un système d'épuration et le règlement de la pénalisation, c'est-à-dire que

$$F_i\,(\Phi_i) = Y_i + V_i + w_i. \tag{7.5}$$

Ce système est hiérarchique. Le Représentant de l'Administration occupe une position privilégiée dans la mesure où il a la possibilité de fixer à l'avance le montant de la pénalisation $c$. Donc, nous ferons l'analyse de son point de vue. Tout d'abord il doit formuler des hypothèses sur le comportement (les objectifs) des autres agents. L'hypothèse suivante est plus ou moins vraisemblable: le choix des commandes dont disposent les entreprises est défini par la condition

$$J_i = V_i + w_i \Rightarrow \min. \tag{7.6}$$

En fait, cette condition exprime la tendance des entreprises à accroître constamment leurs fonds. En d'autres termes, le critère (7.6) équivaut au critère

$$Y_i \Rightarrow \max.$$

Etant donné qu'au moment de la prise de décision les entreprises connaîtront le montant de la pénalisation, leurs stratégies seront définies à partir des conditions

$$V_i + c\pi_i = V_i + cf_i\,(F_i(\Phi_i),\ V_i) \Rightarrow \min,\quad i = 1,\ \ldots,\ N, \tag{7.7}$$

sous les contraintes

$$V_i \geqslant 0,\quad V_i + cf_i\,(F_i\,(\Phi_i),\ V_i) \leqslant F_i\,(\Phi_i),\quad i = 1,\ \ldots,\ N. \tag{7.8}$$

La résolution du problème (7.7), (7.8) permet au Représentant de l'Administration de déterminer les investissements consacrés au perfectionnement des technologies et à l'épuration intérieure. Ils dépendront visiblement aussi bien du volume des fonds $\Phi_i$ que du montant $c$ de la pénalisation. La résolution du problème (7.7), (7.8)

nous donne

$$V_i = \Psi_i\,(c,\ \Phi_i). \tag{7.9}$$

Effectuons un calcul illustratif pour une situation simple à l'extrême. Supposons que

$$P_i = s_i\Phi_i,\quad \pi_i = \frac{l_i P_i}{a_i+V_i} = \frac{l_i s_i \Phi_i}{a_i+V_i}.$$

Alors

$$w_i = \frac{c l_i s_i \Phi_i}{a_i+V_i}.$$

Utilisons ces expressions pour composer la fonction objectif (7.6):

$$J_i = V_i + \frac{c l_i s_i \Phi_i}{a_i+V_i}. \tag{7.10}$$

Supposons par ailleurs que la valeur de $V_i$ qui la minimise est située

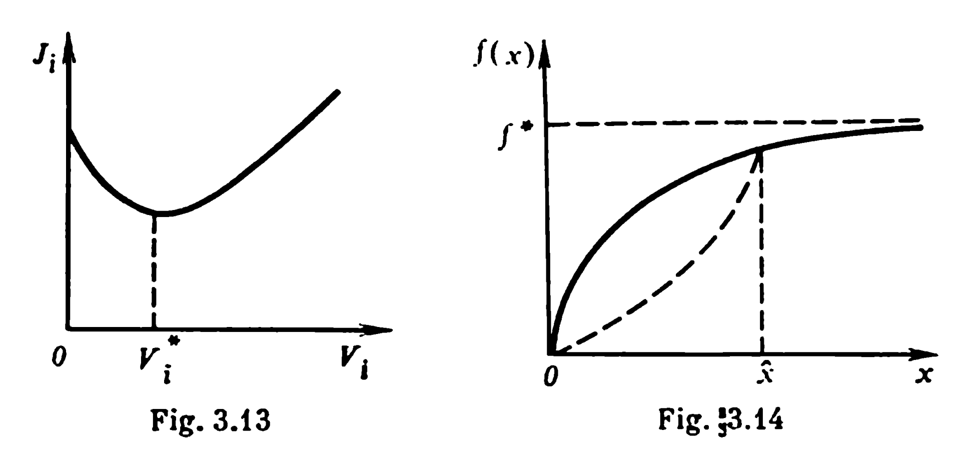



Fig. 3.13

Fig. 3.14

à l'intérieur de l'intervalle déterminé par les conditions (7.8). Le choix de l'entreprise peut être déterminé à partir de la condition

$$\frac{\partial J_i}{\partial V_i} = 0. \tag{7.11}$$

La fonction $J_i$ définie par (7.10) admet un seul minimum (fig. 3.13) $V_i = V_i^*$ qui se déduit de l'équation (7.11):

$$V_i^* = \sqrt{c l_i s_i \Phi_i} - a_i. \tag{7.12}$$

Calculons encore les quantités $\pi_i$ et $w_i$:

$$\pi_i = \frac{1}{\sqrt{c}} \sqrt{l_i s_i \Phi_i}, \tag{7.13}$$

$$w_i = \sqrt{c l_i s_i \Phi_i}. \tag{7.14}$$

Supposons maintenant que l'Administration régionale consacre toutes les ressources tirées des pénalisations à une épuration centralisée

de l'environnement (de l'eau). L'équation caractérisant les variations du niveau de pollution de l'eau peut se mettre alors sous la forme suivante :

$$\dot{x}=\sum \pi_i - f(x) - W\left(\sum w_i\right), \tag{7.15}$$

où $f(x)$ est l'épuration naturelle. En principe, cette fonction est concave et présente un certain seuil de saturation (fig. 3.14) : l'efficacité de l'épuration ne peut excéder une certaine limite $f^*$.

REMARQUE. En réalité la situation est bien plus compliquée et la fonction $f(x)$ présente un caractère d'hystérésis. Supposons que la concentration des matières polluantes a commencé à croître au début ; alors l'intensité de l'autoépuration varie suivant la courbe pleine jusqu'à une valeur $x = \hat{x}$. Ensuite le niveau de pollution commence à baisser. Mais la variation de l'intensité d'épuration ne suivra plus la courbe pleine, elle s'effectuera d'une autre manière, par exemple suivant la courbe en pointillé.

La fonction $W\left(\sum w_i\right)$ représente la quantité de déchets détruits par l'épuration centralisée. Convenons que

$$W=\mu \sum w_i = \mu c \sum \pi_i .$$

Donc, l'équation (7.15) peut s'écrire :

$$\dot{x}=\sum \pi_i (1-\mu c) - f(x). \tag{7.16}$$

Comme chaque entreprise connaîtra au moment de la prise de décision la politique de pénalisation adoptée par l'Administration régionale, cette dernière peut admettre que $V_i$ est déterminée par la formule (7.9) et par suite

$$\pi_i = f_i(P_i, V_i) = f_i(F_i(\Phi_i), \Psi_i(c, \Phi_i)) = f_i^*(\Phi_i, c),$$
$$i = 1, \ldots, N;$$

l'équation (7.16) devient alors

$$\dot{x}=\sum_i f_i^*(\Phi_i, c)(1-\mu c) - f(x). \tag{7.17}$$

Dans le cas particulier considéré (cas pour lequel sont valables les formules (7.12), (7.13) et (7.14)), on aura

$$\dot{x}=\sum_i \sqrt{l_i s_i \Phi_i}\left(\frac{1}{\sqrt{c}} - \mu\sqrt{c}\right) - f(x). \tag{7.18}$$

Analysons maintenant les motivations de l'Administration régionale dans la fixation de la pénalisation. Diverses situations peuvent se présenter et l'Administration régionale peut poursuivre plusieurs buts subjectifs. Discutons quelques versions possibles.

Tout d'abord l'Administration régionale doit viser une non-dégradation de l'environnement, c'est-à-dire que

$$\dot{x}(t) \leqslant 0 \quad \forall t \tag{7.19}$$

ou à la rigueur

$$\dot{x}(t) \Rightarrow \min \quad \forall t, \tag{7.20}$$

si la condition (7.19) est impossible à réaliser.

Les critères (7.19) ou (7.20) ne sont pas les seuls à retenir l'attention de l'Administration. Celle-ci peut encore s'intéresser au développement industriel de la région, qui assure la stabilité du niveau de vie de la population et qui réponde à ses intérêts. Ce critère peut être formulé de diverses manières. On peut par exemple exiger que

$$\frac{d}{dt} \sum \Phi_i > 0 \tag{7.21}$$

ou

$$\frac{d}{dt} \sum \Phi_i \Rightarrow \max. \tag{7.22}$$

Le critère (7.20) peut se mettre sous la forme suivante

$$\sum_i f_i^* (\Phi_i, c)(1-\mu c) - f(x) \Rightarrow \min \tag{7.23}$$

ou pour le cas particulier qui nous occupe

$$I_1(c) = \sum_i \sqrt{l_i s_i \Phi_i} \left(\frac{1}{\sqrt{c}} - \mu \sqrt{c}\right) - f(x) \Rightarrow \min. \tag{7.24}$$

Transformons maintenant le critère (7.22). L'équation (7.1) et l'égalité (7.5) nous permettent d'écrire:

$$\frac{d}{dt} \sum \Phi_i =$$

$$= \sum \{F_i(\Phi_i) - \Psi_i(\Phi_i, c) - c f_i^* (\Phi_i, c) - k\Phi_i\} \Rightarrow \max. \tag{7.25}$$

Pour le cas particulier envisagé, la relation (7.25) s'écrit

$$I_2(c) = \sum \{\bar{s}_i \Phi_i - 2\sqrt{c l_i s_i}\, \Phi_i - a_i\} \Rightarrow \max, \tag{7.26}$$

où $\bar{s}_i = s_i - k_i$.

Nous sommes ainsi conduits au problème à double critère:

$$I_1(c) \Rightarrow \min, \qquad I_2(c) \Rightarrow \max.$$

Considérons maintenant les relations (7.24) et (7.26). Elles sont représentées graphiquement sur la figure 3.15. Appelons $\hat{c}$ la racine de l'équation $I_1(c) = 0$ et $c^*$, celle de l'équation $I_2(c) = 0$. Si

$\hat{c} < c^*$, alors pour tout $c \in ]\hat{c}, c^*[$ seront réalisées les conditions

$$\dot{x} < 0, \quad \frac{d}{dt} \sum \Phi_i > 0,$$

c'est-à-dire que dans ce cas on peut fixer des pénalisations qui d'une part contribueront à réduire le niveau de pollution et de l'autre n'entraveront pas la croissance des fonds. Pour déterminer le montant de $c$, il faut construire l'ensemble de Pareto. Ceci n'est pas compliqué dans notre cas. A partir de l'équation (7.26), on déduit $\sqrt{c}$ comme

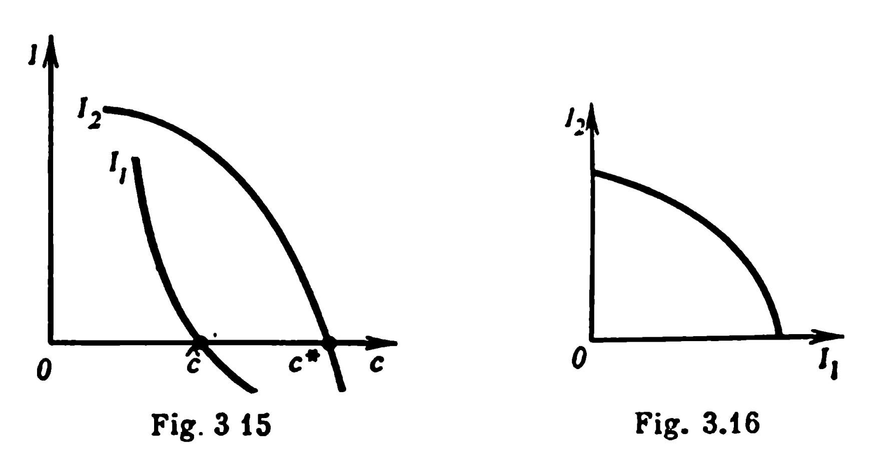



Fig. 3 15 Fig. 3.16

une fonction de $I_2$ $(c)$ et on porte la valeur trouvée dans (7.24). On obtient en définitive la relation

$$I_1 = \varphi(I_2). \tag{7.27}$$

Si l'on connaît les courbes représentées sur la figure 3.15 on peut facilement construire le graphique de l'ensemble de P reto (fig. 3.16). La nature de cet ensemble dépendra naturellement des fonds.

Si $\hat{c} > c^*$, il n'existe pas de domaine de valeurs de $c$ dans lequel on aurait simultanément $\dot{x} < 0$ et $\sum \dot{\Phi}_i > 0$. Dans ce cas, soit nous devons prendre notre parti de la croissance de la pollution si nous voulons que les fond croissent avec le temps, soit nous devons accepter que les fonds fixes soient consacrés à la protection de l'environnement, c'est-à-dire nous accommoder d'une réduction de la production.

Signalons encore un trait curieux de ce problème. Si l'effet de l'auto-épuration n'est pas élevé (par exemple pour un faible niveau de pollution), alors la quantité $\hat{c}$ dépend peu du volume des fonds:

$$\hat{c} \approx 1/\mu.$$

La quantité $c^*$ croît avec les fonds et de plus

$$c^* = O\left(\sum \Phi_i\right).$$

Donc, si au départ $\hat{c} > c^*$, une stratégie payante pour l'Administration régionale est la suivante. Il n'y a pas lieu de se soucier de l'environnement tant que le potentiel industriel de la région est bas. Dès que le volume des fonds atteint une certaine valeur critique que l'on obtient sans peine à partir de la condition

$$\hat{c} = c^*,$$

il faut commencer à pénaliser les entreprises pour la pollution de l'environnement et le montant de cette pénalisation doit croître avec le potentiel industriel.

Nous avons envisagé une situation mettant en jeu deux fonctions objectifs relativement simples, caractérisées par une « optimalité locale »: tous les agents tentent de maximiser leurs intérêts immédiats. Mais la situation est plus naturelle encore dans laquelle les agents planifient leur ligne d'action avec plusieurs années à l'avance. Ainsi, par exemple, l'entreprise peut essayer de maximiser à terme le niveau de ses fonds. Elle est prête dans un premier temps à s'acquitter de la pénalisation pour la pollution de l'environnement et à ne rien investir dans les installations d'épuration quitte à pallier à ces insuffisances au cours du prochain plan. L'Administration régionale peut de même se fixer des objectifs à long terme. Dans ces conditions, l'analyse simple effectuée dans ce paragraphe est insuffisante et l'on doit utiliser des méthodes de simulation.

**REMARQUE.** L'expérimentation de modèles complexes gagne toujours à être précédée de raisonnements semblables à ceux que nous venons de faire : les procédures de simulation doivent toujours comporter une analyse de modèles simplifiés.

## CHAPITRE IV

# MÉTHODES ASYMPTOTIQUES EN ANALYSE DES SYSTÈMES (CAS RÉGULIER)

### § 1. Discussion préliminaire

Au chapitre précédent, j'ai amené le lecteur à l'idée que la simulation, technique de systèmes homme — machine conjuguant l'intellect de l'analyste aux possibilités d'une analyse formelle effectuée sur ordinateur, c'est-à-dire une synthèse de méthodes basées sur l'expérience, l'intuition et le talent et sur une analyse mathématique rigoureuse, est le fer de lance de l'analyse des systèmes. Plus, l'analyse des systèmes est avant tout une utilisation rationnelle des idées et techniques de la simulation. L'analyse des systèmes ne se serait jamais érigée en discipline autonome n'était la contribution des expériences réalisées sur les systèmes de simulation.

Le rôle et la place de la simulation dans les études de systèmes sont aujourd'hui assez bien appréhendés par un grand nombre d'analystes. Mais s ils n'hésitent pas à reconnaître le rôle souverain de l'expérimentation mathématique et des techniques de simulation, ils ramènent par contre souvent le problème tout entier à des questions purement techniques d'organisation des massifs de données, de structure du système opérationnel, de langage d'entrée, etc. Et, bien sûr, ils conviennent tous que le modèle doit refléter assez bien la réalité.

Mais si important qu'ils soient, ces éléments de simulation ne suffisent pas à eux seuls. L'expérimentation sur ordinateur est impossible sans de bonnes méthodes mathématiques, sans une théorie spéciale de l'expérimentation sur ordinateur, l'utilisation directe du modèle n'étant pas toujours aisée et justifiée.

Les difficultés rencontrées par l'expérimentateur d'un modèle sont nombreuses et de nature diverse. Elles ne sont pas toujours d'ordre mathématique. Plus le modèle est compliqué, plus il est difficile de le saturer d'information non contradictoire, d'organiser les calculs, etc. Il existe aussi des difficultés purement mathématiques.

En effet, pour qu'un modèle décrive bien la réalité, il doit nécessairement être assez compliqué. Mais dans ce cas chaque expérience prendra beaucoup de temps. Ce qui rend quasiment impossible la réalisation du nombre nécessaire d'expériences, condition *sine qua non* de toute analyse.

Donc, le premier obstacle que l'analyste trouve sur son chemin est la limitation des ressources. Les calculs sur le modèle doivent être assez économiques. La résolution de ces problèmes est impossible sans la participation de la théorie des schémas aux différences, surtout si le modèle est décrit par des équations aux dérivées partielles. Mais on admettra que le bâtisseur du modèle et de son logiciel possède les fondements du calcul numérique ce qui nous dispensera d'exposer les approximations finidimentionnelles dans cet ouvrage.

Supposons donc (ce que du reste nous ferons toujours dans la suite) que le passage aux équations aux différences s'est déroulé dans les meilleures conditions et que les méthodes de calcul retenues sont les plus économiques.

Nous n'en avons pas pour autant fini avec les difficultés d'ordre mathématique. Il y a encore le problème de la dimension, le problème de l'échelle (différence des temps caractéristiques) des diverses composantes du phénomène étudié, etc.

Tout ceci met l'analyste dans une situation qui semble parfois inextricable. Donc, la simplification du modèle, son remplacement par un autre, plus accessible à l'analyse, est le problème clef de l'analyste. Le cybernéticien R. Ashby qui est un grand spécialiste de la théorie des systèmes a préconisé de traiter cette discipline comme la science de simplifier les systèmes étudiés.

La description de ces possibilités dépasse le cadre de cet ouvrage. On se bornera donc à produire seulement quelques méthodes spéciales permettant non seulement de simplifier le modèle, mais de construire aussi pour certains cas particuliers des algorithmes « rapides ».

Dans ce chapitre et dans les suivants, nous nous attarderons sur la discussion des diverses particularités de la résolution des classes de problèmes de Cauchy qui se posent souvent à l'analyste et qui généralement sont omis dans les cours d'Analyse universitaires.

Nous commencerons donc notre exposé par la discussion du problème de Cauchy. Il existe plusieurs raisons à cela. En général, le problème de Cauchy, c'est-à-dire la détermination d'une trajectoire de phase correspondant à des commandes données, des perturbations données et des conditions initiales données, est l'élément vital de presque toutes les procédures de simulation.

Même lorsque le problème principal de l'analyste consiste à trouver une commande optimale ou lorsqu'on construit l'ensemble permis ou l'ensemble de Pareto, la résolution du problème de Cauchy est la plus importante phase d'algorithmes plus complexes. Nous nous sommes déjà assurés de cela au chapitre deux lorsque nous avons discuté la possibilité de construire les commandes optimales. Donc, l'examen des difficultés soulevées par la résolution du problème de Cauchy nous permet d'avancer à grands pas dans l'appréhension des

écueils qu'il faut surmonter pour organiser les expériences de simulation.

Ainsi, le problème majeur qui se pose à l'analyste a trait à la longue occupation de l'ordinateur pour résoudre le problème de Cauchy. Cela tient à plusieurs raisons dont les plus significatives sont :

a) La complexité des opérateurs décrivant le processus physique étudié (leur structure qui reflète des relations non linéaires complexes) et le grand nombre d'opérations logiques.

b) L'existence de processus oscillatoires intérieurs qui imposent un petit pas d'intégration.

c) L'existence de domaines de brusque variation d'un groupe de variables (couches frontières intérieures).

c) La dimension élevée du problème.

Chacune de ces circonstances qui freinent la conduite de l'expérimentation numérique a été d'une manière ou d'une autre étudiée en analyse ou dans des recherches appliquées. Et un usage éclairé des méthodes respectives est un gage de réussite dans de nombreux cas. Donc, toute expérience doit être précédée d'une analyse des systèmes et de leur traitement préliminaire. L'étude préliminaire du problème commence par l'utilisation de diverses procédures non formelles.

L'analyse de toute situation plus ou moins complexe passe toujours par une restriction successive de l'ensemble des variantes, par l'élimination des variantes non concurrentielles. Cette voie répond visiblement aux particularités purement physiologiques du cerveau humain, aux particularités du raisonnement humain et peut-être même aux particularités de la création *). Sa réalisation s'appuie souvent sur des procédures intuitives, non formelles (et non formalisées). L'analyse séquentielle relève néanmoins dans de nombreux cas de méthodes mathématiques rigoureuses.

Ainsi, les idées de l'analyse séquentielle qui ont été inspirées déjà par les travaux de Markov ont connu un important rayonnement. Elles ont conduit en fin de compte à l'élaboration d'un puissant instrument, tel que la programmation dynamique et ses généralisations du genre méthode des branches et frontières. Nous avons pris connaissance de certaines d'entre elles au chapitre II. Les autres seront examinées ultérieurement.

L'application du « principe de Rodin » qui revêt généralement un caractère non formel localise le problème et ne laisse qu'un nombre peu important de variantes qui sont analysées par d'autres méthodes.

---

*) Rappelons à ce propos la réponse de Rodin à la question de savoir comment il travaillait sur ses œuvres : « C'est très simple, je prends un bloc de marbre et j'enlève tout ce qui est superflu. » Pour cette raison le principe de restriction successive des alternatives est parfois appelé principe de Rodin.

L'autre approche se base sur la sélection de mouvements « voisins », sur la construction de diverses théories des perturbations qui permettraient de comparer des classes de décisions voisines selon tel ou tel critère. En analyse des systèmes, ces méthodes ont conduit à l'apparition des concepts d'« algorithmes rapides », « systèmes tampons », « calculs vérificatifs », etc.

Le schéma général de ces méthodes d'analyse est plus ou moins simple. Etant donné que le système de modèles initial est généralement assez complexe (en tout cas assez pour rendre impossible la réalisation d'un grand nombre d'expériences), on le remplace par un modèle plus simple dont l'analyse (c'est-à-dire la résolution du problème de Cauchy, des problèmes d'optimisation et autres) n'implique pas beaucoup de temps machine et peut être conduite dans des délais raisonnables. Les algorithmes d'analyse des modèles simplifiés s'appellent *algorithmes rapides*.

Les algorithmes rapides jouent un rôle éminent en analyse des systèmes. Ils permettent tout d'abord de mettre au rébut d'autres variantes encore. En principe, les algorithmes rapides aidant, on peut former l'ensemble des modèles retenus pour l'expérimentation. Les calculs effectués à l'aide du modèle accepté sont dits *vérificatifs*.

REMARQUE. Entre les calculs rapides et vérificatifs, il y a la même relation qu'entre les calculs d'ingénieur sur le projet d'un avion et ses essais de vol. Mais si crédibles que soient les calculs sur les modèles simplifiés, ils doivent s'achever par des essais sur « nature ». Mais en dépit de toute la perfection possible, les essais du prototype sont trop coûteux pour servir au choix du système.

Ainsi, les algorithmes rapides sont très importants en analyse des systèmes et sont la clef de voûte des calculs numériques.

Parfois il est absolument indispensable de savoir préciser les résultats de ces calculs que nous conviendrons d'appeler dans la suite *calculs estimatifs*, de mettre en évidence les erreurs possibles ou la perte de précision qui sont inhérentes aux algorithmes rapides. A cet effet ont été mises au point des théories spéciales des perturbations qui font l'objet des chapitres IV, V et VI. Les systèmes de calcul effectués à l'aide de la théorie des perturbations s'appellent *systèmes tampons*. Les calculs divers de l'expérimentation sont représentés sur la figure 4.1. Les flèches doubles indiquent qu'en principe le processus de simulation est un système itératif spécialement organisé.

REMARQUE. La création de tout gros projet revêt un caractère hiérarchique. On discute d'abord le principe général, ensuite on commence à projeter les divers éléments du système et à chaque fois on retrouve le schéma de la figure 4.1 sous une forme ou une autre.

Le système tampon est au même titre que le système de calculs estimatifs un important élément des procédures de simulation. Il permet non seulement de préciser les calculs estimatifs, mais d'éliminer les variantes non concurrentielles. La principale fonction du

système tampon est de comparer les calculs vérificatifs, c'est-à-dire les calculs effectués sur le modèle accepté, aux calculs réalisés à l'aide des algorithmes rapides.

Dans ce chapitre, nous entamerons l'étude des méthodes de construction des systèmes tampons. Aux difficultés soulevées par les algorithmes de calcul viennent se greffer des difficultés de principe. Lorsqu'on fait appel à des schémas simplifiés d'analyse il reste

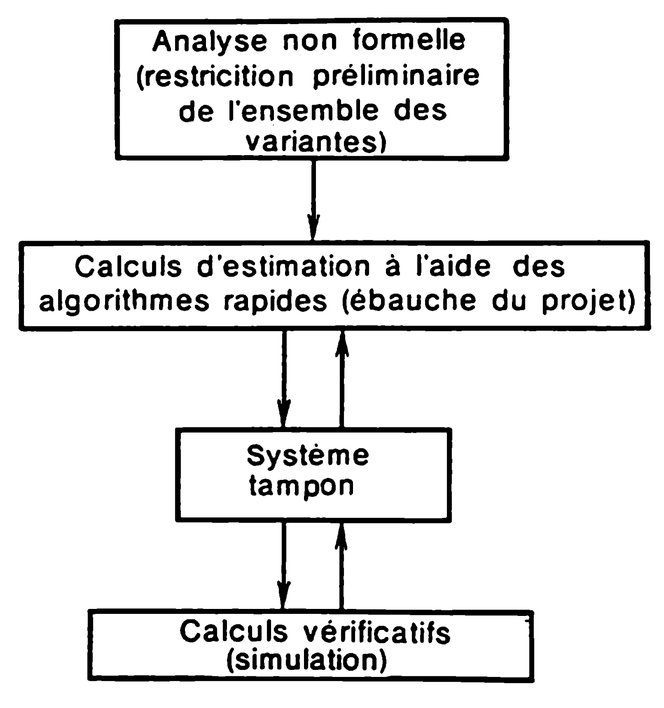



Fig. 4.1

toujours le problème de la compréhension des conditions dont la réalisation nous garantit que l'élimination de certains facteurs secondaires n'occultent pas les qualités qui ont suscité l'analyse.

Donc, l'analyse préliminaire qui précède l'expérimentation pose de nombreux problèmes de principe. Il faut mettre en place les techniques nécessaires d'analyse approchée et indiquer les voies qui permettront de construire les estimations et d'établir la qualité de l'analyse approchée.

Malgré toute la diversité des phénomènes étudiés et des méthodes de construction des modèles simplifiés et des systèmes tampons, il se dessine une tendance à l'unification des méthodes respectives d'analyse. Cette entreprise est menée à bien à l'aide de la théorie générale du petit paramètre. Cette théorie dont les bases ont été jetées au siècle dernier déjà est destinée au problème suivant. Soit donné un système d'équations contenant un paramètre:

$$F(x, \dot{x}, t, \varepsilon) = 0. \qquad (1.1)$$

où $F$ est une fonction vectorielle de même dimension que le vecteur $x$ et supposons que pour $\varepsilon = 0$, une intégrale de ce système est connue, soit

$$x^0 = x^0 (t, c), \tag{1.2}$$

où $c$ est le vecteur des constantes arbitraires. Le système d'équations $F(x, \dot{x}, t, 0) = 0$ sera appelé *système générateur*, et la solution (1.2) correspondant à une valeur particulière de $c$, *solution génératrice.*

Quels liens existent-ils entre une solution génératrice et les solutions du système (1.1) pour des valeurs assez petites du paramètre $\varepsilon$? Les solutions génératrices conservent-elles les principales propriétés de la solution du système (1.1)? Comment construire (si cela est possible) les itérations qui permettront de trouver avec n'importe quelle précision la solution du système (1.1) en s'aidant de l'information sur une solution génératrice? Comment construire une solution approchée en s'appuyant sur une solution génératrice et estimer la précision de la solution approchée? Telles sont les principales questions auxquelles les systèmes tampons sont appelés à répondre.

Toutes les méthodes qui préciseront la solution génératrice seront appelées méthodes de la théorie des perturbations. Ce chapitre est consacré à l'exposé de certains aspects de cette théorie et à des exemples illustrant son usage en analyse des systèmes.

## § 2. Théorie classique de Poincaré

Dans cet ouvrage on se bornera à l'étude des systèmes susceptibles d'être résolus par rapport aux dérivées supérieures, c'est-à-dire aux systèmes de la forme

$$\dot{x} = f(x, t, \varepsilon), \tag{2.1}$$

et citons certains faits qui nous permettront d'introduire des simplifications qui faciliteront non seulement l'analyse qualitative mais aussi les expériences de simulation. Le comportement de la solution de l'équation (2.1) pour $\varepsilon \to 0$ s'appelle *comportement asymptotique* de cette solution et l'analyse de ses propriétés, *analyse asymptotique.* Donc, le problème fondamental de la théorie des perturbations est l'analyse asymptotique.

Le premier et l'un des plus importants résultats de l'analyse asymptotique est le célèbre théorème de Poincaré. Ce théorème se rapporte au cas où la fonction $f$ est une fonction analytique de la variable indépendante $x$ et du paramètre $\varepsilon$, et établit que la solution est une fonction analytique du paramètre.

Posons le problème de Cauchy pour le système (2.1): déterminer une fonction $x(t, \varepsilon)$ vérifiant le système (2.1) et les conditions initiales

$$x(0) = x_0. \tag{2.2}$$

Pour que le problème (2.1), (2.2) ait un sens, il suffit d'exiger que le second membre de l'équation (2.1) vérifie des conditions garantissant l'existence d'une solution locale, par exemple la condition de Lipschitz. On considérera de pair avec l'équation (2.1) l'équation génératrice

$$\dot{z} = f(z, t, 0). \tag{2.3}$$

On posera le même problème de Cauchy (2.2) pour ces équations.

Dans l'équation (2.1) faisons la substitution $x = z + y$, où $z(t)$ est une fonction connue. Le vecteur $y$ satisfera l'équation

$$\dot{y} = f(z + y, t, \varepsilon) - f(z, t, 0) \tag{2.4}$$

et les conditions initiales nulles

$$y(0) = 0. \tag{2.5}$$

Le second membre de l'équation (2.4) étant une fonction analytique de $y$ et $\varepsilon$, on peut le développer en série de Taylor suivant ces variables en les admettant petites en module. Ceci nous permet de mettre l'équation (2.4) sous la forme du développement suivant les puissances de $\varepsilon$:

$$\dot{y} = Ay + \varepsilon\left(\frac{\partial f}{\partial \varepsilon}\right)_0 + B(y, \varepsilon, t), \tag{2.6}$$

où $A$ est une matrice carrée des dérivées partielles premières: $A = \left(\frac{\partial f^i}{\partial x^j}\right)_0$; $f^i$ et $x^j$ les composantes respectives des vecteurs $f$ et $x$; $\left(\frac{\partial f}{\partial \varepsilon}\right)_0$ le vecteur (matrice colonne) de composantes $\left(\frac{\partial f^i}{\partial \varepsilon}\right)_0$, $B(y, \varepsilon, t)$, l'ensemble des termes d'ordre supérieur: le développement de la fonction $B(y, \varepsilon, t)$ commence par les carrés des arguments $y$ et $\varepsilon$; la matrice $A$ et le vecteur $\frac{\partial f}{\partial \varepsilon}$ sont calculés pour $x = z$, $\varepsilon = 0$.

Si l'on suppose que la solution du problème de Cauchy de l'équation génératrice est connue, on peut admettre que les matrices $A$ et $\left(\frac{\partial f}{\partial \varepsilon}\right)$ sont des fonctions connues du temps.

Cherchons la solution de l'équation (2.6) sous la forme de la série

$$y = \sum_{i=1}^{\infty} y_i \varepsilon^i. \tag{2.7}$$

En portant cette série dans l'équation (2.6) et en identifiant les coefficients des mêmes puissances du paramètre $\varepsilon$, on obtient le système

d'équations suivant pour la détermination des fonctions $y_i$:

$$\begin{aligned}\dot{y}_1 &= Ay_1 + D_1,\\ \dot{y}_2 &= Ay_2 + D_2,\\ &\cdots\cdots\\ \dot{y}_k &= Ay_k + D_k.\end{aligned} \tag{2.8}$$

Dans ces équations $D_1 = \left(\frac{\partial f}{\partial \varepsilon}\right)_0$ est une fonction vectorielle du temps connue. La fonction vectorielle $D_2$ contient les termes quadratiques du développement de la fonction $B(y, \varepsilon, t)$ suivant $y$ et $\varepsilon$, c'est-à-dire qu'elle ne contient que la fonction $y_1$ et pas les fonctions $y_i$ avec $i > 1$. En effet, la fonction $B(y, \varepsilon, t)$ peut être mise sous la forme

$$B(y, \varepsilon, t) = B_{00}y^2 + B_{01}y\varepsilon + B_{11}\varepsilon^2 + \ldots \tag{2.9}$$

En portant la série (2.7) dans cette expression, on obtient

$$B(y, \varepsilon, t) = \varepsilon^2 (B_{00}y_1^2 + B_{01}y_1 + B_{11}) + \varepsilon^3 (\ldots) + \ldots$$

Il est évident que chaque fonction $D_k$ ne dépend que des fonctions $y_i$ telles que $i < k$.

Donc, si l'on résout successivement les équations (2.8), il faut traiter les fonctions $D_k$ comme des fonctions connues du temps. Les fonctions $y_i$ vérifient les conditions initiales nulles:

$$y_i(0) = 0. \tag{2.10}$$

Ainsi, la détermination des coefficients du développement de la solution $y(t)$, c'est-à-dire des fonctions $y_i(t)$, se ramène à la résolution successive de problèmes de Cauchy pour le système (2.8).

Nous sommes maintenant en mesure de formuler le théorème de Poincaré.

THÉORÈME DE POINCARÉ. I. *Si l'on connaît l'intégrale générale de l'équation génératrice* (2.3), *on peut trouver la solution du système d'équations* (2.8) *par une différentiation et une intégration.*

II. *La solution du système d'équations* (2.1) *est une fonction analytique du paramètre* $\varepsilon$, *c'est-à-dire que les séries* (2.7) *convergent pour des valeurs de* $\varepsilon$ *assez petites en module et par suite sont des intégrales des équations* (2.1), *développées suivant les puissances de* $\varepsilon$.

On conviendra d'appeler solution formelle une série vérifiant formellement le système d'équations différentielles; cette série sera appelée solution si elle converge dans un domaine de valeurs du paramètre. Donc, la première partie du théorème donne une appréciation de la structure de l'algorithme de construction de la solution formelle. La deuxième partie affirme que la solution formelle (2.7) est une solution.

Bornons-nous à la démonstration de la première proposition du théorème. Nous glisserons sur la démonstration de la deuxième proposition, car elle est assez laborieuse. On peut la trouver par exemple dans [31].

Soit

$$z = F(t, C),$$

où $C = (C^1, \ldots, C^n)$ est une constante arbitraire (un vecteur de dimension $n$), l'intégrale générale de l'équation génératrice (2.3). Cela signifie tout d'abord que la fonction $F$ vérifie l'équation (2.3):

$$\frac{dF}{dt} = f[F(t, C), t, 0] \tag{2.11}$$

pour toute valeur de la constante $C$. Désignons par $\xi_i$ le vecteur

$$\xi_i = \frac{\partial F(t, C)}{\partial C^i}. \tag{2.12}$$

Calculons

$$\frac{d\xi_i}{dt} = \frac{d}{dt}\frac{\partial F}{\partial C^i} = \frac{\partial}{\partial C^i}\frac{dF}{dt}.$$

Utilisons l'identité (2.11):

$$\frac{d\xi_i}{dt} = \frac{\partial}{\partial C^i}\{f(F(t, C), t, 0)\} = \frac{\partial f}{\partial F}\frac{\partial F}{\partial C^i} = \frac{\partial f}{\partial F}\xi_i,$$

où $\partial f/\partial F$ est la matrice carrée

$$\frac{\partial f}{\partial F} = \left(\frac{\partial f^i}{\partial F^j}\right) = \left(\frac{\partial f^i}{\partial z^j}\right) = \left(\frac{\partial f^i}{\partial x^j}\right)_0.$$

De là il est évident que

$$\frac{\partial f}{\partial F} = A.$$

Donc, la fonction vectorielle $\xi_i$ est solution du système d'équations différentielles linéaires

$$\dot{\xi}_i = A\xi_i$$

pour tout $i$. Le système d'équations

$$\dot{u} = Au$$

s'appelle équations aux variations pour le système (2.1): c'est un système d'équations différentielles linéaires à coefficients variables. Il n'existe aucune recette pour l'intégration des équations à coefficients variables. Mais les équations aux variations sont douées d'une remarquable propriété que nous venons tout juste d'établir. Formulons cette propriété sous la forme du lemme suivant.

15—01523

LEMME. *Si l'on connaît l'intégrale générale de l'équation génératrice, on peut expliciter les solutions particulières des équations aux variations par une différentiation conformément à la formule* (2.12).

Ainsi, la formule (2.12) définit le système de solutions fondamentales des équations aux variations. Maintenant on peut déduire la solution du problème de Cauchy (2.8), (2.10) par des quadratures en se servant de la méthode de variation des constantes arbitraires. Citons ces calculs. Le système d'équations (2.8) s'écrit

$$\dot{y} = Ay + D, \tag{2.13}$$

où $D$ est une fonction connue du temps. Cherchons la solution sous la forme

$$y = Y\alpha, \tag{2.14}$$

où $\alpha$ est un vecteur inconnu, $Y$, la matrice des solutions fondamentales des équations aux variations: $Y = \{\xi_i^j\}$. Donc,

$$\frac{dY}{dt} = AY. \tag{2.15}$$

En dérivant (2.14) (compte tenu de (2.15)) et en portant le résultat obtenu dans (2.13) on trouve

$$Y\dot{\alpha} = D, \tag{2.16}$$

ou

$$\dot{\alpha} = Y^{-1}D,$$

d'où finalement

$$y = Y(t)\left\{\int_0^t Y^{-1}(\tau)\,D(\tau)\,d\tau + C^*\right\},$$

où le vecteur $C^*$ est une nouvelle constante d'intégration.

Pour que les conditions initiales (2.10) soient satisfaites, il faut que la constante $C^*$ soit nulle. Donc

$$y = \int_0^t Y(t)\,Y^{-1}(\tau)\,D(\tau)\,d\tau. \tag{2.17}$$

La matrice $Y(t)\,Y^{-1}(\tau)$ s'appelle *matrice de Green.*

La première partie du théorème de Poincaré est prouvée *in extenso*, puisque l'expression (2.17) a été déduite par une différentiation et des intégrations.

Le théorème de Poincaré a inspiré diverses simplifications et dans de nombreux cas peut servir de base à la construction d'algorithmes rapides, de systèmes tampons et pour la *décomposition du problème*, c'est-à-dire le remplacement d'un problème de dimension élevée par des sous-problèmes de dimension plus petite. A ces problèmes se rapportent les *systèmes faiblement liés.*

Considérons le système

$$\dot{x} = f(x, \varepsilon y), \qquad \dot{y} = \varphi(y, \varepsilon x), \tag{2.18}$$

où $x$ et $y$ sont des vecteurs de dimensions respectives $n$ et $m$. Si $\varepsilon = 0$ ce système se décompose en deux systèmes indépendants:

$$\dot{x} = f(x, 0), \tag{2.19}$$

$$\dot{y} = \varphi(y, 0), \tag{2.20}$$

dont la construction des trajectoires se ramène à la résolution successive de deux problèmes de Cauchy indépendants:

$$x(0) = x_0, \qquad y(0) = y_0. \tag{2.21}$$

La résolution successive des problèmes (2.19), (2.21) et (2.20), (2.21) dont la dimension est inférieure à celle du problème initial est bien moins laborieuse que celle du problème initial (2.18).

La méthode de résolution du problème (2.19), (2.21) (ou du problème (2.20), (2.21)) peut jouer le rôle d'algorithme rapide. Une fois en possession de cette solution, que l'on désignera par $x^0$ et $y^0$, on peut grâce au théorème de Poincaré estimer la précision de cette approximation (par rapport à la solution exacte du système (2.18)). Si l'on désigne les solutions exactes par $x(t)$, $y(t)$, on aura de toute évidence

$$x = x^0 + O(\varepsilon), \qquad y = y^0 + O(\varepsilon), \tag{2.22}$$

c'est-à-dire que l'erreur sera de l'ordre de $\varepsilon$. Ce qui veut dire que cette erreur tendra vers 0 avec $\varepsilon$. Il est en principe presque impossible d'obtenir des estimations plus exactes avec de telles théories asymptotiques.

Pour obtenir une solution plus exacte, on peut se servir de la procédure développée plus haut ; à cet effet, il faut représenter les fonctions cherchées sous la forme des séries

$$\begin{aligned} x &= x^0 + \varepsilon x_1 + \varepsilon^2 x_2 + \ldots, \\ y &= y^0 + \varepsilon y_1 + \varepsilon^2 y_2 + \ldots \end{aligned} \tag{2.23}$$

On obtient les solutions précisées en se bornant à un certain nombre de termes. Dans de nombreux problèmes pratiques, il suffit de se limiter aux deux premiers termes de ces développements. Dans ce cas la théorie des perturbations se ramène à l'intégration des équations

$$\begin{aligned} \dot{x}_1 &= \left(\frac{\partial f}{\partial x}\right)_0 x_1 + \left(\frac{\partial f}{\partial \varepsilon}\right)_0 y^0, \\ \dot{y}_1 &= \left(\frac{\partial \varphi}{\partial y}\right)_0 y_1 + \left(\frac{\partial \varphi}{\partial \varepsilon}\right)_0 x^0. \end{aligned}$$

Le calcul de l'approximation suivante passe par la linéarisation des équations initiales et ensuite par la résolution du problème de Cauchy relatif aux équations linéaires obtenues. Cette procédure a sa raison d'être si les méthodes d'analyse sont « manuelles ». Mais, si l'on a l'intention de se servir d'un ordinateur, il faut avoir présent à l'esprit que des équations non linéaires écrites sous une forme plus compacte sont plus commodes à analyser numériquement que les équations linéaires. Par ailleurs, la procédure de linéarisation est souvent assez laborieuse et implique le calcul des dérivées, opération qui se traduit souvent par une perte de précision. Donc, pour construire la théorie des perturbations, c'est-à-dire des systèmes de calculs permettant de préciser les approximations $x^0$ et $y^0$ acquises pour $\varepsilon = 0$, il faut procéder autrement et résoudre les problèmes initiaux de Cauchy mais pour les systèmes

$$\dot{x} = f(x, \varepsilon y^0), \qquad \dot{y} = \varphi(y, \varepsilon x^0). \tag{2.24}$$

Donc, le système de recalculs que nous avons appelé système tampon se ramène, si les conditions du théorème de Poincaré sont réunies, à l'intégration numérique de deux systèmes d'équations (2.24).

Les variables ne présentent pas toutes de l'intérêt dans les systèmes de dimension élevée. Certaines d'entre elles (par exemple $y(t)$) peuvent être des degrés de liberté « parasites ». Si une telle situation se présente, il suffit d'inclure seulement le premier système (2.24) dans le système tampon.

## § 3. Quelques exemples

a) *Fonctionnement de populations liées.* Le modèle classique de Volterra, dit modèle prédateur — proie, est décrit par le système d'équations différentielles

$$\dot{x}_1 = f(x_2)\, x_1, \qquad \dot{x}_2 = \varphi(x_1)\, x_2. \tag{3.1}$$

Ce système repose sur un modèle élémentaire de la dynamique des populations. Désignons par $z(t)$ le nombre d'individus d'une population. Alors la forme la plus élémentaire de description des variations du nombre d'individus de cette population est

$$\dot{z} = \gamma z, \tag{3.2}$$

où $\gamma = \alpha - \beta$ ; $\alpha$ s'appelle taux de natalité, $\beta$, taux de mortalité. Si ces taux sont constants, nous avons affaire à une population à croissance exponentielle. Les équations de type (3.2) n'ont un sens que si le nombre d'individus est assez élevé et la fonction $z(t)$, continue.

Considérons l'interaction de deux populations dans le cadre de cette représentation élémentaire de la dynamique des populations. L'une d'elles sera composée de $x_1$ ($t$) prédateurs. Pour cette population la quantité $\gamma$ dépendra du nombre d'individus $x_2$ ($t$) de l'autre population, qui seront les proies. On admet que le nombre de prédateurs ne définit que le taux de natalité, le taux de mortalité étant constant et défini par la mortalité naturelle, c'est-à-dire que $f(x_2) = f_1(x_2) - a_{11}$. Le coefficient $f_1(x_2)$ est une fonction concave de $x_2$ (fig. 4.2). La fonction $f_1$ admet l'approximation linéaire $f_1(x_2) = a_{12}x_2$ pour des valeurs pas trop élevées de $x_2$. De façon analogue,

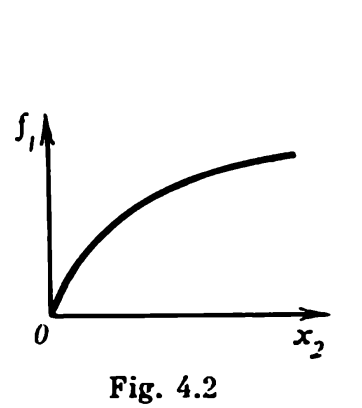


Fig. 4.2

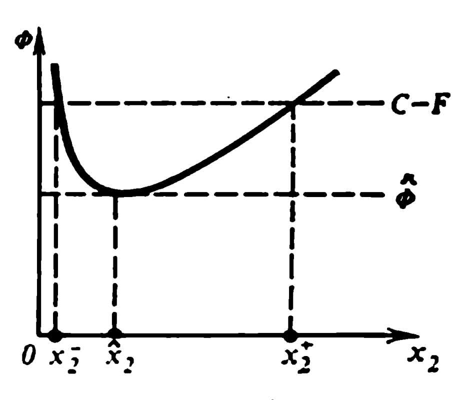


Fig. 4.3

la dynamique du nombre de proies est entièrement définie par le nombre de prédateurs. Ceci étant, on admet qu'aucune proie ne meurt de mort naturelle et qu'elle sert à la nutrition des prédateurs. S'agissant du coefficient de natalité, il ne dépend pas du nombre de prédateurs et est constant pour la population considérée. En d'autres termes, $\varphi(x_1) = a_{22} - \varphi_1(x_1)$, et de plus $\varphi_1(x_1)$ est aussi une fonction concave de $x_1$ et qui est justiciable aussi de l'approximation linéaire $\varphi_1(x_1) = a_{21}x_1$. En regroupant ces résultats, on est conduit au système d'équations

$$\dot{x}_1 = -a_{11}x_1 + a_{12}x_1x_2, \qquad \dot{x}_2 = -a_{21}x_1x_2 + a_{22}x_2, \qquad (3.3)$$

connu sous le nom de *modèle de Volterra.*

Nous ne discuterons pas la signification physique ou, plus exactement, biologique de ce modèle. Elle est de toute évidence assez conventionnelle. Néanmoins elle décrit la structure de relations trophiques qui reflètent de nombreuses particularités qualitatives des processus réels trophiques. Les modèles de type (3.3) ont joué un rôle immense dans le devenir de la théorie mathématique des macrosystèmes biologiques. Ils constituent les modèles les plus élémentaires d'interaction de populations.

Il est évident que les équations (3.3) admettent la solution stationnaire

$$\hat{x}_1 = a_{22}/a_{21}, \qquad \hat{x}_2 = a_{11}/a_{12}. \qquad (3.4)$$

Si $x_1 = \hat{x}_1$, $x_2 = \hat{x}_2$ à un instant $t = t_0$, alors ces égalités restent valables pour $t > t_0$.

D'autre part, la solution triviale $x_1 \equiv x_2 \equiv 0$ est la seule solution stationnaire du système (3.3).

Les variables $x_1$ et $x_2$ sont positives de par leur signification physique, c'est-à-dire que le système (3.3) sera étudié dans le premier quadrant.

Si les états initiaux ne seront pas confondus avec les quantités (3.4), alors les trajectoires du système (3.3) seront des fonctions périodiques du temps. Prouvons cette proposition.

Signalons tout d'abord que le système (3.3) admet une intégrale première. Pour le prouver, calculons

$$\frac{dx_2}{dx_1} = \frac{(a_{22} - a_{21}x_1)\,x_2}{(a_{12}x_2 - a_{11})\,x_1}.$$

Les variables sont séparables dans cette équation et l'on trouve sans peine

$$\Phi(x_2) = a_{12}x_2 - a_{11}\ln x_2 = \\ = C - (a_{21}x_1 - a_{22}\ln x_1) \equiv C - F(x_1), \qquad (3.5)$$

où $C$ est la constante d'intégration.

Considérons la fonction $\Phi(x_2)$ ; elle est représentée graphiquement sur la figure 4.3 ; elle présente un minimum $\hat{\Phi}$ au point $\hat{x}_2$. Si $C - F < \hat{\Phi}$, alors l'équation

$$\Phi(x_2) = C - F \qquad (3.6)$$

ne présente pas de racines réelles. Si $C - F > \hat{\Phi}$, alors l'équation (3.6) admet deux racines strictement positives.

La fonction $F(x_1)$ est de nature analogue. Désignons maintenant par $\hat{C}$ la valeur de la constante arbitraire pour laquelle

$$\hat{\Phi} = \hat{C} - F(\hat{x}_1).$$

La fonction $F(x_1)$ étant strictement positive et $x_1 = \hat{x}_1$ étant un point de minimum, le système (3.3) n'admet pas de trajectoires réelles pour tout $C < \hat{C}$. A la valeur $C = \hat{C}$ du plan de phase $(x_1, x_2)$ correspondra un seul point de coordonnées $x_1 = \hat{x}_1$, $x_2 = \hat{x}_2$. Pour $C > \hat{C}$ on a l'inégalité

$$C - F(\hat{x}_1) \geqslant \hat{\Phi}; \qquad (3.7)$$

à toute valeur de $x_1$ d'un voisinage de $x_1 = \hat{x}_1$ correspondront deux valeurs strictement positives de la variable $x_2$ (fig. 4.4). Mais si la valeur de $x_1$ est extérieure à l'intervalle défini par la condition

(3.7), alors aux valeurs de $x_1$ ne correspondront plus des valeurs réelles de $x_2$. Les extrémités de l'intervalle auquel sont associées les trajectoires réelles, se déterminent à partir de l'équation

$$C - F(x_1) = \hat{\Phi},$$

qui admet deux racines strictement positives: $x_1^+$ et $x_1^-$ (cf. fig. 4.4). Ainsi, à toute valeur de $C > \hat{C}$ sera associée une trajectoire fermée du plan de phase.

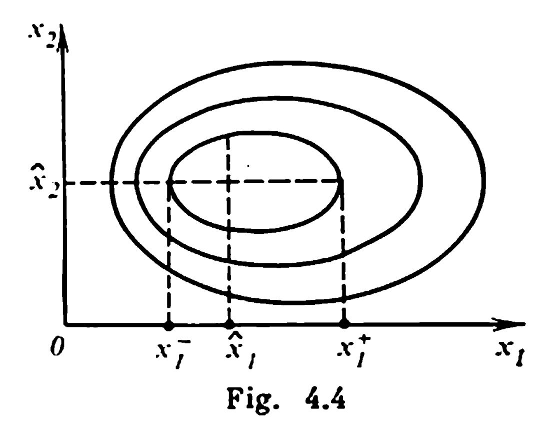



Fig. 4.4

Le plan de phase est représenté sur la figure 4.4. Le point (0, 0) appartient aussi à la famille de solutions, mais par lui il ne passe aucune trajectoire de phase *)

REMARQUE. Le système (3.3) admet des solutions périodiques car c'est un système de Liapounov: ses seconds membres sont des fonctions analytiques de ses variables et il admet une intégrale analytique (3.5). Pour plus de détails voir [7].

Pour la suite de l'exposé nous aurons besoin de la seule région du plan de phase située au voisinage du point $(\hat{x}_1, \hat{x}_2)$. Pour décrire les trajectoires dans ce voisinage on se contentera d'une approximation linéaire. On peut de toute évidence la déduire de l'intégrale (3.5), mais il est plus simple de la calculer directement. Posons

$$x_1 = \hat{x}_1 + \xi_1, \qquad x_2 = \hat{x}_2 + \xi_2. \tag{3.8}$$

En portant (3.8) dans (3.3) et en négligeant les quantités du second ordre de petitesse par rapport à $\xi_1$ et $\xi_2$, on obtient

$$\dot{\xi}_1 = \frac{a_{12}a_{22}}{a_{21}}\xi_2, \quad \dot{\xi}_2 = -\frac{a_{21}a_{11}}{a_{12}}\xi_1 \tag{3.9}$$

ou

$$\ddot{\xi}_1 + \omega^2\xi_1 = 0, \tag{3.10}$$

*) La structure du modèle et son étude détaillée ont été données au début du siècle déjà par V. Volterra. Cf. [14] pour plus de détails.

où

$$\omega^2 = a_{11}a_{22}. \tag{3.11}$$

L'équation (3.10) donne

$$\dot{\xi}_1 = \xi_{1,0} \cos(\omega t + \alpha),$$

où $\xi_{1,0}$ et $\alpha$ sont des constantes arbitraires.

Le système (3.9) admet l'intégrale première

$$\xi_1^2 \frac{a_{21}}{a_{12}a_{22}} + \xi_2^2 \frac{a_{12}}{a_{21}a_{11}} = C. \tag{3.12}$$

Les formules (3.10), (3.11) et (3.12) nous donnent une description qualitative et quantitative correcte du fonctionnement du système de populations considéré dans le cas où son état initial est proche du point $(\hat{x}_1, \hat{x}_2)$.

Supposons maintenant que nous avons affaire à un autre système de populations prédateurs — proies dont le fonctionnement est régi par les équations

$$\dot{y}_1 = -b_{11}y_1 + b_{12}y_1y_2, \qquad \dot{y}_2 = -b_{21}y_1y_2 + b_{22}y_2. \tag{3.13}$$

Dans le système (3.3) nous avons admis que les taux étaient strictement positifs. Dans ce système, le taux de natalité $b_{22}$ des proies sera supposé strictement négatif, soit $b_{22} < 0$. Cette situation peut se présenter lorsque la population $y_2$ fait l'objet d'une exploitation, par exemple la pêche est proportionnelle au nombre de poissons et ce coefficient de proportionnalité est supérieur au taux de natalité, ou lorsque cette population est frappée par une épidémie, etc. Dans ces conditions, il est évident que même en l'absence de prédateurs (c'est-à-dire lorsque $y_1 \equiv 0$) la population $y_2$ décroîtra indéfiniment. Cette situation prévaudra *a fortiori* en présence de prédateurs. Donc

$$\lim_{t\to\infty} y_2(t) = 0.$$

Mais, dans ces conditions, le nombre de prédateurs décroîtra indéfiniment, puisque à partir d'une certaine date $y_2$ sera si petit ($y_2 < < y_2^*$) que $b_{12}y_2 < b_{11}$ et le second membre de l'équation pour $\dot{y}_1$ sera strictement inférieur à $-(b_{11} - b_{12}y_2^*)\, y_1$. Donc, dans ces conditions toute solution du système (3.13) décroîtra indéfiniment.

Supposons maintenant que les deux systèmes (3.3) et (3.13) sont reliés par exemple de la manière suivante:

$$\begin{aligned}
\dot{x}_1 &= -a_{11}x_1 + a_{12}x_1x_2 + \varepsilon d_1x_1y_2,\\
\dot{x}_2 &= -a_{21}x_1x_2 + a_{22}x_2 - \varepsilon c_1y_1x_2,\\
\dot{y}_1 &= -b_{11}y_1 + b_{12}y_1y_2 + \varepsilon d_2x_2y_1,\\
\dot{y}_2 &= -b_{21}y_1y_2 + b_{22}y_2 - \varepsilon c_2x_1y_2,
\end{aligned} \tag{3.14}$$

autrement dit le système (3.14) décrit la situation dans laquelle les prédateurs se nourrissent des individus des deux populations.

Si le paramètre $\varepsilon$ est petit, alors pour étudier la solution génératrice on peut le prendre égal à 0 et prendre pour solution approchée la solution des problèmes précédents.

Le théorème de Poincaré est valable puisque les seconds membres du système d'équations (3.14) sont des fonctions analytiques du paramètre $\varepsilon$. Ceci nous conduit aux conclusions suivantes.

1) Quels que soient les états initiaux $x_{1,0}$ et $x_{2,0}$ du premier système, les variations des populations des prédateurs et des proies suivront une loi proche d'oscillations périodiques: l'accroissement du nombre de prédateurs entraînera une diminution du nombre de proies, c'est-à-dire une restriction de la base nutritive des prédateurs, ce qui aura à la longue pour conséquence une réduction du nombre de ces derniers. Or, ceci se traduira par contre-coup par un accroissement du nombre de proies, c'est-à-dire par un élargissement de la base nutritive des prédateurs, donc par un éventuel accroissement de leur nombre, et ainsi de suite.

2) S'agissant du deuxième système de populations, ses deux composantes diminueront progressivement et tendront indéfiniment vers zéro pour $t \to \infty$.

En vertu des estimations du paragraphe précédent, ces conclusions seront d'autant plus exactes que $\varepsilon$ sera petit. En d'autres termes, si l'on remplace la résolution numérique du problème de Cauchy relatif au système (3.14) par la résolution de deux sous-problèmes de Cauchy de dimension deux fois moindre, on obtient des estimations qualitatives de la trajectoire de phase dont l'erreur sera d'autant plus petite que $\varepsilon$ sera petit. Ce résultat appelle des commentaires.

En effet, en prouvant le théorème de Poincaré on s'est essentiellement servi du fait que l'intervalle de temps est fixe et fini. Lorsque l'intervalle de temps croît, les estimations se relâchent constamment et on ne peut plus mettre en évidence certaines caractéristiques qualitatives importantes des processus étudiés à l'aide de la théorie développée. Enfin, le théorème de Poincaré s'énonce en termes de théories asymptotiques: il décrit le comportement des solutions pour $\varepsilon \to 0$. En réalité, les paramètres sont toujours petits, mais finis, et il se pose la question des limites d'utilisation de la théorie, du domaine de valeurs des paramètres dans lequel ils restent assez petits pour que les conclusions qualitatives établies à l'aide de l'analyse des solutions génératrices soient valables.

Si $b_{22} < 0$, alors on a vu que $\lim\limits_{t \to \infty} y_2 = 0$. Considérons ce cas limite. Le système (3.14) s'écrit maintenant

$$\begin{aligned} \dot{x}_1 &= -a_{11}x_1 + a_{12}x_1x_2, \\ \dot{x}_2 &= -a_{21}x_1x_2 + a_{22}x_2 - \varepsilon c_1 y_1 x_2, \\ \dot{y}_1 &= -b_{11}y_1 + \varepsilon d_2 x_2 y_1. \end{aligned} \tag{3.15}$$

Le système (3.15) n'admet plus de solutions stationnaires (sauf pour le cas où $\frac{a_{11}}{a_{12}} = \frac{b_{11}}{\varepsilon d_2}$ dans lesquelles les quantités $x_1$, $y_1$ et $x_2$ soient différentes de 0. Pour étudier le comportement des solutions du système (3.15) pour $t \to \infty$ il faut procéder à une analyse spéciale qui peut être conduite de la manière suivante. Mettons la première et la troisième équation sous la forme suivante:

$$\dot{x}_1/x_1 = -a_{11} + a_{12}x_2, \qquad \dot{y}_1/y_1 = -b_{11} + \varepsilon d_2 x_2;$$

d'où

$$\varepsilon d_2 d\,(\ln x_1) - a_{12} d\,(\ln y_1) = (b_{11}a_{12} - \varepsilon a_{11}d_2)\,dt,$$

ou

$$x_1^{\varepsilon d_2}/y_1^{a_{12}} = C \exp\,(b_{11}a_{12} - \varepsilon a_{11}d_2)\,t;$$

alors

$$y_1 = C^* x_1^{\varepsilon d_2/a_{12}} \exp\left(-\frac{b_{11}a_{12} - \varepsilon a_{11}d_2}{a_{12}}\,t\right). \tag{3.16}$$

De la formule (3.16) il s'ensuit immédiatement que si $b_{11}a_{12} - \varepsilon a_{11}d_2 > 0$, alors $\lim_{t\to\infty} y_1 = 0$, c'est-à-dire qu'on obtient un résultat conforme à la conclusion qualitative fournie par l'analyse de la solution génératrice.

Si l'on a l'inégalité inverse, c'est-à-dire $b_{11}a_{12} - \varepsilon a_{11}d_2 < 0$, alors la situation change: la première population de prédateurs disparaît et laisse la place à l'autre population. Donc les conclusions déduites à l'aide de l'analyse effectuée grâce au théorème de Poincaré ne sont valables que si

$$\varepsilon < \varepsilon^* = \frac{b_{11}a_{12}}{a_{11}d_2}. \tag{3.17}$$

Signalons que le théorème de Poincaré est valable non seulement lorsque $\varepsilon \to 0$, mais aussi lorsque $\varepsilon$ est assez petit. Toutefois, il est en général assez difficile de localiser cette frontière et d'établir des conditions de type (3.17).

REMARQUE. La formule (3.17) montre que dans ce problème le rôle du petit paramètre est tenu non pas par un seul paramètre, mais par une combinaison de paramètres, plus exactement, par le rapport des valeurs de ces paramètres.

b) *Système à élément oscillatoire.* Supposons que nous avons un système écologique dont l'état est décrit par un vecteur $z$ de dimension $n$ et par un système prédateur — proie de très petites périodes d'oscillations propres $T = 2\pi/\omega$. Ceci signifie que si le système prédateur — proie était placé dans des conditions stationnaires, alors la période $2\pi/\omega$, où $\omega$ est donné par la formule (3.11), serait petite. La notion de petitesse est toujours plus ou moins relative. Dans le cas

considéré, elle signifie que si nous éliminons du système écologique le couple prédateur – proie, alors les autres variables varieront lentement: le temps caractéristique de variation du vecteur $z$ est de beaucoup supérieur à $T$. En d'autres termes, nous étudions le comportement d'un système oscillatoire prédateur — proie sur un fond variant lentement.

Nous admettrons que les activités vitales des populations se répercutent sur l'environnement et que le milieu ambiant influe sur le fonctionnement du système prédateur — proie.

Mettons les équations du système étudié sous la forme suivante:

$$\begin{aligned} \dot{x} &= -a_{11}x + a_{12}xy, \\ \dot{y} &= -a_{21}xy + a_{22}(z)\, y, \\ \dot{z} &= f(z, x, y). \end{aligned} \tag{3.18}$$

On supposera que les quantités $a_{11}$, $a_{12}$ et $a_{21}$ sont des paramètres constants. Cela signifie qu'elles ne dépendront ni du temps, ni des valeurs des autres variables du problème. S'agissant du taux de natalité $a_{22}$ de la population qui sert de nutrition au prédateur, on admettra qu'il dépend de l'état du milieu.

Si $z$ = const, alors le nombre de prédateurs et de proies oscille au voisinage de la position d'équilibre, qui comme on le sait est définie par les formules

$$x = \hat{x} = a_{22}/a_{21}, \qquad y = \hat{y} = a_{11}/a_{12}.$$

Bornons-nous au cas où $x$ et $y$ ne s'écartent pas fortement des valeurs stationnaires $\hat{x}$ et $\hat{y}$. Considérons le changement de variables

$$x = \hat{x} + \xi_1, \qquad y = \hat{y} + \xi_2. \tag{3.19}$$

Linéarisons les deux dernières équations du système (3.18) par rapport à $\xi_1$ et $\xi_2$. On obtient le système d'équations suivant en $\xi_1$, $\xi_2$ et $z$ (cf. formules (3.9)):

$$\begin{aligned} \dot{z} &= f(z, \hat{x} + \xi_1, \hat{y} + \xi_2), \\ \dot{\xi}_1 &= \frac{a_{12}a_{22}(z)}{a_{21}} \xi_2, \\ \dot{\xi}_2 &= -\frac{a_{21}a_{11}}{a_{12}} \xi_1. \end{aligned} \tag{3.20}$$

Les équations (3.20) décrivent l'évolution du système (3.18) au voisinage de l'état stationnaire (3.4).

Le système (3.20) ne contient pas encore le petit paramètre. La condition que la période $T$ est petite (ou que la fréquence $\omega$ est gran-

de) n'est pas encore répercutée dans le système. Donc, dans les équations (3.20) faisons le changement de variables suivant (appelé changement de Van der Pol):

$$\xi_1 = c\cos\varphi, \quad \dot{\xi}_1 = -\sqrt{a_{11}a_{12}}\, c\sin\varphi, \tag{3.21}$$

où $c$ et $\varphi$ qui seront appelés par la suite *amplitude* et *phase*, sont les nouvelles variables. En comparant l'expression (3.21) de $\dot{\xi}_1$ avec la deuxième équation du système (3.20), on peut exprimer $\xi_2$ en fonction des nouvelles variables $c$ et $\varphi$:

$$\xi_2 = -\frac{a_{21}}{a_{12}}\sqrt{\frac{a_{11}}{a_{22}}}\, c\sin\varphi. \tag{3.22}$$

Dérivons la première égalité (3.21) et égalons-la à la seconde. On obtient la condition de compatibilité des transformations (3.21):

$$\dot{c}\cos\varphi - c\dot{\varphi}\sin\varphi = -\sqrt{a_{11}a_{22}}\, c\sin\varphi. \tag{3.23}$$

On reconnaît une équation du premier ordre reliant deux fonctions inconnues $c$ et $\varphi$. Pour obtenir une deuxième équation reliant ces fonctions, on dérive l'égalité (3.22) et on l'égale au second membre de la troisième équation (3.20). Des calculs immédiats nous donnent

$$\sqrt{\frac{a_{11}}{a_{22}}}\,\dot{c}\sin\varphi + \sqrt{\frac{a_{11}}{a_{22}}}\, c\dot{\varphi}\cos\varphi =$$
$$= a_{11}c\cos\varphi + \frac{1}{2}\sqrt{\frac{a_{11}}{a_{22}}}\,\dot{a}_{22}c\sin\varphi. \tag{3.24}$$

Le système d'équations (3.23), (3.24) sera traité comme un système d'équations différentielles par rapport à $\dot{c}$ et $\dot{\varphi}$. En le résolvant par rapport à ces quantités et en posant $\omega^2 = a_{11}a_{12}$, on obtient

$$\dot{c} = \frac{\dot{a}_{22}}{2a_{22}}\sin^2\varphi, \quad \dot{\varphi} = \omega + \frac{\dot{a}_{22}}{ca_{22}}\cos\varphi\sin\varphi.$$

En remplaçant enfin les quantités $\xi_1$ et $\xi_2$ de la première équation du système (3.20) par leurs expressions en fonctions de $c$ et $\varphi$, on ramène ce système à la forme

$$\begin{aligned} \dot{z} &= F(z, c, \varphi), \\ \dot{c} &= \Psi_1(z, c, \varphi)\sin^2\varphi, \\ \dot{\varphi} &= \omega + \Psi_2(z, c, \varphi)\cos\varphi\sin\varphi, \end{aligned} \tag{3.25}$$

où les nouvelles notations ont une signification évidente.

Le système (3.25) est entièrement équivalent au système (3.20), mais il possède les deux caractéristiques suivantes: premièrement, les fonctions $F$, $\Psi_1$ et $\Psi_2$ sont des fonctions $2\pi$-périodiques de $\varphi$.

Deuxièmement, ce système contient explicitement la quantité $\omega$ qui définit la période du cycle de regénération du maillon volterrien de notre système écologique. La dernière caractéristique est particulièrement importante dans le cas qui nous intéresse: lorsque $\omega$ est assez grand, le temps caractéristique de fonctionnement des populations est sensiblement inférieur à celui des variations des paramètres du milieu ambiant.

Posons $\omega = \omega_0 \Omega(z)$, où $\omega_0$ est la valeur initiale de $\omega$ et introduisons un nouveau temps $\tau$ « rapide »:

$$\tau = \omega_0 t. \tag{3.26}$$

Admettons que la quantité $\omega_0$ est grande et introduisons le petit paramètre $\varepsilon = 1/\omega_0$. Le système (3.25) s'écrit alors

$$\begin{aligned} dz/d\tau &= \varepsilon F(z, c, \varphi), \\ dc/d\tau &= \varepsilon \Psi_1(z, c, \varphi) \sin^2 \varphi, \\ d\varphi/d\tau &= \Omega(z) + \varepsilon \Psi_2(z, c, \varphi) \cos \varphi \sin \varphi. \end{aligned} \tag{3.27}$$

Le système (3.27) contient explicitement le petit paramètre et de plus les seconds membres des équations dépendent analytiquement de ce paramètre, donc l'on peut formellement utiliser la méthode de Poincaré de représentation des solutions par les séries:

$$z = z_0 + \varepsilon z_1 + \ldots, \quad c = c_0 + \varepsilon c_1 + \ldots, \quad \varphi = \varphi_0 + \varepsilon \varphi_1 + \ldots \tag{3.28}$$

Si le paramètre $\varepsilon$ est assez petit, alors les séries (3.28) convergent et représentent exactement la solution du système d'équations (3.27). Voyons maintenant s'il est possible de déterminer par cette méthode la solution du problème, c'est-à-dire voyons dans quelle mesure on peut appliquer les séries (3.28) pour obtenir les réponses aux questions concernant l'évolution de l'environnement et l'influence sur ce dernier des caractéristiques de l'activité du système donné de deux populations.

En portant les séries (3.28) dans le système (3.27) et en identifiant les coefficients des mêmes puissances de $\varepsilon$, on obtient des systèmes d'équations pour la détermination de $z_i$, $c_i$ et $\varphi_i$. La première approximation nous donne

$$z = z_0 = \text{const}, \qquad c = c_0 = \text{const}, \qquad \varphi = \Omega(z_0)\tau = \omega t. \tag{3.29}$$

Cette approximation ne présente pratiquement pas d'intérêt pour l'analyste qui étudie l'évolution graduelle de l'environnement. En effet, les expressions (3.29) ne décrivent que le caractère des variations d'un paramètre du système de Volterra au voisinage de la position d'équilibre pour des conditions constantes du milieu am-

biant :

$$x=\frac{a_{22}(z_0)}{a_{21}}+c\cos\omega t,$$
$$y=\frac{a_{11}}{a_{12}}-\frac{a_{21}}{a_{12}}\sqrt{\frac{a_{11}}{a_{22}(z_0)}}\,c\sin\omega t, \qquad (3.30)$$

et ce mouvement est connu *a priori* de par la position du problème. Donc, il nous faut étudier au moins la deuxième approximation.

La détermination de l'approximation suivante se ramène à l'intégration d'un système linéaire d'équations dont les seconds membres contiennent des fonctions périodiques de $t$, de période $2\pi/\omega$. Ce sont des fonctions rapidement oscillatoires dont le caractère des variations définit la valeur du pas de l'intégration numérique. Aussi pour étudier les variations lentes des paramètres du milieu ambiant faut-il prendre plusieurs pas temporels. Non seulement le calcul sera laborieux, mais il sera entaché d'erreurs à cause du nombre élevé de pas.

La procédure décrite est peu efficace encore pour la raison suivante. Les séries (3.28) convergent pour tout $t$ fini, mais lorsque $t$ croît, la précision de l'approximation effectuée avec les séries de Poincaré diminue.

Par conséquent, même s'il est possible de se servir directement de la méthode de Poincaré, il est peu probable que celle-ci soit utile pour l'étude du problème considéré, problème dont l'objet principal est l'étude de l'évolution du milieu ambiant à long terme. Pour résoudre ce problème, il faut sensiblement modifier la méthode du petit paramètre. Nous verrons plus bas que cela est possible.

Le problème que nous venons de décrire et sur lequel nous reviendrons encore dans ce chapitre appartient à une classe assez vaste de problèmes de technique et d'économie qui étudient des processus oscillatoires rapides sur un fond variant lentement. A ces problèmes se rapportent par exemple les problèmes de la dynamique d'un engin cosmique pénétrant dans l'atmosphère terrestre (cf. [71]).

c) *Solutions périodiques de l'équation de Duffing.* Lorsqu'on étudie des processus de nature physique diverse, il est souvent nécessaire d'établir l'éventualité de l'existence et des singularités des solutions périodiques. Les formes les plus simples de l'état d'un système sont les formes stationnaires. Viennent ensuite les formes périodiques. Depuis Kepler, les mouvements périodiques ont tenu une place très importante dans les recherches des mathématiciens, physiciens, astronomes. Les mouvements périodiques et les mouvements proches d'eux jouent un grand rôle en physique et en technique, où ils sont souvent les formes déterminantes de l'état. On sait maintenant que de nombreux processus biologiques, écologiques et économiques suivent une loi périodique ou quasi périodique. On peut sans crainte d'exagérer dire que pratiquement tous les changements évolutifs relèvent de la répétition (ou presque) d'un grand nombre de processus,

d'états et de phénomènes. Pour cette raison, l'étude des mouvements périodiques sera conduite avec l'arsenal des méthodes mathématiques utilisées en analyse des systèmes. Nous reviendrons à maintes reprises sur ces questions. Dans ce numéro on se propose d'examiner une seule question : la possibilité d'appliquer la méthode de Poincaré à l'étude des mouvements périodiques.

Considérons le cas élémentaire de l'*équation de Duffing*:

$$\ddot{x} + x - x^3 = 0. \qquad (3.31)$$

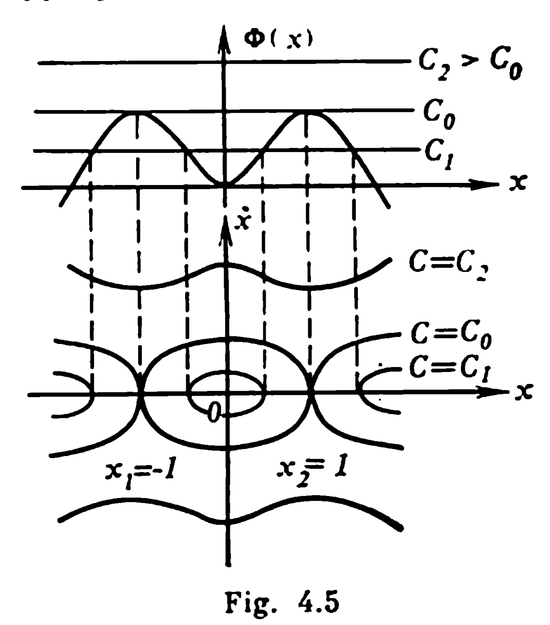


Fig. 4.5

L'étude systématique des oscillations non linéaires a commencé au milieu du siècle dernier. L'équation de Duffing était probablement le premier (ou l'un des premiers) exemple d'équation essentiellement non linéaire à permettre l'établissement de nombreuses caractéristiques qualitatives des mouvements oscillatoires de systèmes non linéaires qui différenciaient ces derniers des systèmes linéaires.

L'équation (3.31) peut être intégrée sous la forme explicite au moyen de fonctions elliptiques et étudiée intégralement par les méthodes analytiques. On se propose une autre démarche. L'existence d'une solution périodique de l'équation (3.31) peut être établie assez f cilement par des constructions géométriques à l'aide d'une intégrale première de cette équation. Pour obtenir une intégrale première, multiplions les deux membres de l'équation (3.31) par $\dot{x}$ et mettons-la sous la forme

$$\ddot{x}\dot{x} + x\dot{x} - x^3\dot{x} \equiv \frac{1}{2}\frac{d\dot{x}^2}{dt} + \frac{1}{2}\frac{dx^2}{dt} - \frac{1}{4}\frac{dx^4}{dt} = 0,$$

d où

$$\dot{x} = \pm\sqrt{C - \Phi(x)}, \qquad (3.32)$$

où $\Phi(x) = x^2 - x^4/2$, et $C$ est une constante arbitraire.

Pour construire les trajectoires de l'équation de Duffing, traçons le graphique de $\Phi(x)$ (fig. 4.5) et posons $C = C_1 < C_0$, où $C_0 = \max\{x^2 - x^4/2\} = 1/2$. A toute valeur de $|x| < 1$ seront associées, en vertu de (3.32), deux valeurs de la variable $\dot{x}$. A l'ensemble de ces valeurs correspondent deux courbes symétriques qui constituent une courbe ellipsoïdale fermée comme l'indique la partie inférieure de la figure 4.5.

Dans les domaines $x > 1$ et $x < -1$ les trajectoires ne sont pas fermées. Pour $C = C_0$ la trajectoire est une séparatrice.

Par conséquent, si l'écart initial par rapport à la position d'équilibre $x = 0$ est petit, alors il engendre des trajectoires fermées qui sont des mouvements périodiques.

Observons que si l'on étudie l'équation

$$\ddot{x} + x + x^3 = 0,$$

alors on constatera que toutes les trajectoires seront des courbes fermées bornées, c'est-à-dire que la solution ne contiendra pas de mouvements illimités.

Ces raisonnements n'épuisent pas l'étude. Il faut encore trouver la période d'oscillation et la relation $x(t)$. Pour cela intégrons l'équation (3.31). Ceci est immédiat, puisque les variables sont séparables dans (3.31) et la solution peut être représentée par des fonctions elliptiques (cf. [7]). On obtient en définitive des formules assez volumineuses. Mais il est possible d'obtenir des expressions approchées relativement simples. Essayons d'utiliser pour cela la formule de Poincaré en considérant le mouvement avec une amplitude assez petite. L'équation (3.31) ne contient pas explicitement le petit paramètre. Le rôle de ce dernier est tenu par l'état initial du système, état qui s'écarte peu de la position d'équilibre. Ce fait peut être utilisé de plusieurs manières: en particulier, on peut grâce à lui introduire un paramètre. Posons $x = \varepsilon y$. L'équation (3.31) devient alors

$$\ddot{y} + y - \varepsilon^2 y^3 = 0. \tag{3.33}$$

Prenons pour condition initiale

$$\dot{y}(0) = 0, \qquad y(0) = \alpha \tag{3.34}$$

L'équation (3.31) ne contient pas de termes dépendant explicitement du temps. Elle est donc invariante par le groupe de translations

$$t \to t + c. \tag{3.35}$$

Donc, nous pouvons toujours choisir la constante $c$ de telle sorte que soient satisfaites les conditions (3.34). La solution vérifiant ces conditions sera de la forme $y = y(t, \alpha)$, c'est-à-dire contiendra une constante arbitraire: or, en vertu de (3.35), on peut la mettre sous la forme

$$y = y(t + c, \alpha). \tag{3.36}$$

La fonction (3.36) contient déjà deux constantes arbitraires, c'est-à-dire est l'intégrale générale.

Dans la suite nous nous servirons à plusieurs reprises de ces circonstances. Mais dans le cas présent la situation est plus simple: on sait *a priori* que toute solution de l'équation (3.31) est périodique

si seulement $\varepsilon$ est assez petit. Donc, pour voir si l'on peut utiliser la méthode de Poincaré au calcul des solutions périodiques, il suffit de considérer une solution quelconque de cette équation et notamment celle définie par les conditions (3.34).

Essayons de trouver la solution du problème de Cauchy (3.33) (3.34) sous forme de la série

$$y = y_0 + \varepsilon y_1 + \varepsilon^2 y_2 + \dots \tag{3.37}$$

On obtient les équations suivantes pour les termes de cette série:

$$\ddot{y}_0 + y_0 = 0, \qquad \ddot{y}_1 + y_1 = 0, \; \ddot{y}_2 + y_2 = y_0^3, \; \dots \tag{3.38}$$

De la première équation on déduit $y_0 = A \cos t + B \sin t$, où $A$ et $B$ sont des constantes arbitraires.

Imposons aux valeurs initiales les conditions (3.34): alors $B = 0$, $A = \alpha$ et l'on obtient

$$y_0 = \alpha \cos t.$$

Les conditions (3.34) étant déjà vérifiées de par le choix des constantes arbitraires $A$ et $B$, les autres fonctions inconnues devront satisfaire les conditions initiales nulles:

$$\begin{aligned} y_1(0) &= y_2(0) = y_3(0) = \dots = 0, \\ \dot{y}_1(0) &= \dot{y}_2(0) = \dot{y}_3(0) = \dots = 0. \end{aligned} \tag{3.39}$$

La deuxième équation (3.38) étant homogène, en utilisant les conditions (3.39), on trouve

$$y_1(t) \equiv 0.$$

Pour $y_2(t)$ on obtient l'équation

$$\ddot{y}_2 + y_2 = \alpha^3 \cos^3 t,$$

or $\cos^3 t = \frac{3}{4} \cos t + \frac{1}{4} \cos 3t$, donc

$$\ddot{y}_2 + y_2 = \frac{3}{4} \alpha^3 \cos t + \frac{1}{4} \alpha^3 \cos 3t. \tag{3.40}$$

Cherchons une solution particulière de l'équation (3.40) sous la forme

$$\tilde{y}_2 = C_1 t \sin t + C_2 \cos 3t$$

En portant l'expression de $\tilde{y}_2$ dans l'équation (3.40) et en identifiant les coefficients des fonctions trigonométriques de même multiplicité,

on trouve

$$\tilde{y}_2 = \frac{3}{8}\alpha^3 t \sin t - \frac{\alpha^3}{32}\cos 3t.$$

Mettons la solution générale de l'équation (3.40) sous la forme

$$y_2 = \tilde{y}_2 + A_2 \sin t + B_2 \cos t$$

et choisissons les constantes $A_2$ et $B_2$ de façon à satisfaire les conditions (3.39). Il est immédiat de voir que $A_2 = 0$, $B_2 = \alpha^3/32$. Donc, la solution de l'équation (3.40) sera

$$y_2 = \frac{3\alpha^3}{8} t \sin t - \frac{\alpha^3}{32}(\cos 3t - \cos t). \qquad (3.41)$$

Des calculs analogues nous donnent

$$y_3 \equiv 0, \qquad (3.42)$$

$$y_4 = a_1 t \sin t + a_2 t^2 \cos t + a_3 \cos t + a_4 \cos 3t + a_5 \cos 5t,$$

où la détermination des coefficients inconnus ne pose pas de problèmes.

Les formules (3.41) et (3.42) montrent que tous les termes de la série (3.37), à l'exception du premier, seront soit nuls, soit séculaires, c'est-à-dire croissant avec le temps. La série ainsi obtenue convergera sur tout intervalle de temps fini en vertu du théorème de Poincaré. Donc, les solutions périodiques de l'équation de Duffing peuvent être approchées avec n'importe quelle précision par la méthode de Poincaré par une série composée de termes qui ne sont pas des fonctions périodiques du temps. Certes, cette représentation n'est pas très commode. On peut s'en servir pour étudier par exemple la dépendance de la période d'oscillation par rapport à l'amplitude ainsi que d'autres caractéristiques importantes du processus oscillatoire envisagé. Or, comme la période de la solution sera différente de $2\pi$, qui est la période des termes de la série, il sera très difficile (voire impossible) de déterminer la période même approximativement. Donc le principal écueil dans l'utilisation de cette représentation sera constitué par le fait que toute portion de la série utilisée ne sera pas une fonction périodique du temps.

Enfin, la précision de l'approximation se détériorera continuellement avec le temps et à mesure qu'on étudiera le processus il faudra faire intervenir un plus grand nombre de termes de la série (3.37).

Toutes ces difficultés relèvent du fait que la période $T$ de la solution dépend des conditions initiales et la solution génératrice doit en quelque sorte en tenir compte.

Donc, la méthode du petit paramètre de Poincaré, qui est un puissant instrument d'étude approchée, est loin d'être universelle et nécessite de sensibles modifications pour devenir apte à la réso-

lution de nombreux problèmes d'applications contenant de petits paramètres.

Dans les paragraphes suivants, nous nous attarderons sur des variantes de la méthode du petit paramètre qui permettront de construire les solutions des deux derniers des trois problèmes envisagés dans ce paragraphe.

## § 4. Méthode de Poincaré de calcul des solutions auto-oscillatoires et périodiques des systèmes quasi linéaires

a) *Schéma de la méthode.* La détermination des solutions périodiques des équations différentielles non linéaires est un problème autonome et important. Notre ambition n'est pas de donner un exposé tant soit peu exhaustif des théories en vigueur: nous voulons montrer certaines modifications de la théorie du petit paramètre, susceptibles d'être une source d'idées dans l'étude des systèmes complexes. La mise en évidence des composantes périodiques ou presque périodiques du processus, leur étude et, ce qui est encore plus important, leur radiation de toute étude ultérieure sont des points forts de l'analyse des systèmes. Malheureusement à ce jour on ne dispose encore d'aucune méthode universelle et il est important que l'analyste se représente la genèse des idées et leur évolution graduelle. Pour cela il suffit probablement de poser les principaux jalons.

Dans ce paragraphe, on se limitera donc à la seule étude de la célèbre méthode proposée par Poincaré au siècle dernier, méthode qui fait intervenir des raisonnements ayant servi plus tard de point de départ à la plupart des autres théories.

Considérons une équation quasi linéaire générale du second ordre

$$\ddot{x} + \lambda^2 x = \varepsilon F(\dot{x}, x). \tag{4.1}$$

Lorsque $\varepsilon = 0$, l'équation (4.1) décrit les oscillations d'un pendule mathématique de fréquence $\omega(0) = \lambda$ et de période $T_0 = 2\pi/\omega(0)$. Pour $\varepsilon \neq 0$, la période de la solution de l'équation (4.1) doit dépendre du paramètre $\varepsilon$: $T = T(\varepsilon) = 2\pi/\omega(\varepsilon)$, et de plus $\lim\limits_{\varepsilon\to 0} \omega(\varepsilon) = \omega(0) = \lambda$ en vertu de la dépendance continue par rapport au paramètre. On peut donc poser

$$\omega(\varepsilon) = \frac{\lambda}{1 + g_1\varepsilon + g_2\varepsilon^2 + \ldots} \tag{4.2}$$

et faire le changement de variables

$$t = \frac{\tau}{\lambda}(1 + g_1\varepsilon + g_2\varepsilon^2 + \ldots). \tag{4.3}$$

A la valeur $t = T$ est associée $\tau = 2\pi$, c'est-à-dire que la période de la solution cherchée par rapport à la nouvelle variable ne dé-

pendra plus de $\varepsilon$ et sera égale à $2\pi$. Les nombres $g_i$ sont *a priori* inconnus et doivent être déterminés dans le cadre de la construction de la solution. Mettons l'équation (4.1) sous la forme suivante, compte tenu du changement de variables (4.3):

$$\frac{d^2x}{d\tau^2}+x(1+g_1\varepsilon+g_2\varepsilon^2+\dots)^2=$$
$$=\varepsilon F\left(x,\ \frac{dx}{d\tau}\frac{\lambda}{1+g_1\varepsilon+g_2\varepsilon^2+\dots}\right)\frac{(1+g_1\varepsilon+g_2\varepsilon^2+\dots)^2}{\lambda^2}. \qquad (4.4)$$

Les équations (4.1) et (4.4) ne contenant pas $t$, l'équation (4.4) sera invariante par la transformation $t\to t+c$ et, comme nous l'avons déjà vu, pour l'étudier intégralement il suffit seulement de considérer le problème de Cauchy suivant:

$$t=0: \qquad x(0)=x_0(\varepsilon), \qquad dx/d\tau=0. \qquad (4.5)$$

Signalons que la quantité $x_0(\varepsilon)$ est généralement inconnue *a priori* et elle ne peut être arbitrairement donnée. En effet, dans les situations typiques, les régimes périodiques soit n'existent pas du tout, soit sont des cas exceptionnels qui sont définis par des conditions initiales spéciales de la forme (4.5) dont la détermination est le principal objet d'étude.

Nous sommes ainsi conduits au problème de la détermination du nombre $x_0(\varepsilon)$ et de la solution périodique de l'équation (4.4) engendrée par cet état initial. Et nous savons *a priori* que la période de cette solution est égale à $2\pi$.

La substitution (4.3) est remarquable par le fait que tout en respectant la dépendance analytique des seconds membres de l'équation par rapport au paramètre, c'est-à-dire qu'elle ne viole pas les conditions d'applicabilité du théorème de Poincaré, elle permet de représenter la solution périodique (si elle existe) de l'équation (4.4) sous forme d'une série dont chaque terme est une fonction périodique du temps de période fixe $2\pi$.

En vertu de ce qui vient d'être dit, on cherchera la solution périodique de l'équation (4.4) sous forme d'une série de $\varepsilon$:

$$x=\sum_{k=0}^{\infty}\varepsilon^k x^{(k)}(\tau). \qquad (4.6)$$

En portant la série (4.6) dans l'équation (4.4) et en identifiant les coefficients des mêmes puissances de $\varepsilon$, on obtient pour les fonctions $x^{(i)}$ les équations suivantes:

$$\frac{d^2x^{(0)}}{d\tau^2}+x^{(0)}=0,$$
$$\frac{d^2x^{(1)}}{d\tau^2}+x^{(1)}=\frac{1}{\lambda^2}F\left(x^{(0)},\ \frac{dx^{(0)}}{d\tau}\lambda\right)-2g_1x^{(0)}, \qquad (4.7)$$
$$\dots\dots\dots\dots\dots$$

L'intégration du système (4.7) fera apparaître des constantes arbitraires. Nous devons choisir ces constantes et les inconnues $g_1, g_2, \ldots$ de telle sorte que les solutions des équations soient, premièrement, des fonctions périodiques de $\tau$ de période $2\pi$ et, deuxièmement, que la série (4.6) vérifie les conditions initiales (4.5).

Ceci constitue la méthode de Poincaré de détermination des solutions périodiques. Si les solutions périodiques existent, cette méthode permet de les trouver sous forme de séries dont tous les termes sont des fonctions périodiques de même période $2\pi$.

La première équation du système (4.7) nous donne la solution qui vérifie les conditions (4.5):

$$x^{(0)} = c \cos \tau, \tag{4.8}$$

où $c = x(0) = x_0$. En portant (4.8) dans (4.7), on trouve

$$\frac{d^2 x^{(1)}}{d\tau^2} + x^{(1)} = \frac{1}{\lambda^2} F(c \cos \tau, \; -\lambda c \sin \tau) - 2 g_1 c \cos \tau. \tag{4.9}$$

Pour que l'équation (4.9) admette des solutions périodiques en $\tau$ de période $2\pi$, il est nécessaire et suffisant que son second membre soit orthogonal à $\cos \tau$ et $\sin \tau$. Ces conditions nous donnent deux équations pour la détermination des constantes inconnues $c$ et $g_1$:

$$J(c) = \int_0^{2\pi} F(c \cos \tau, \; -c\lambda \sin \tau) \sin \tau \, d\tau = 0, \tag{4.10}$$

$$g_1(c) = \frac{1}{2\pi c \lambda^2} \int_0^{2\pi} F(c \cos \tau, \; -c\lambda \sin \tau) \cos \tau \, d\tau. \tag{4.11}$$

La première de ces équations est une équation transcendante en $c$, amplitude de la solution génératrice. Cette équation peut ne pas admettre de solution. Ce sera notamment le cas où l'équation initiale n'admet pas de solution périodique, par exemple, lorsque la fonction $F$ est dissipative. Soit $F(x, \dot{x}) = -k\dot{x}$. L'équation (4.10) s'écrit alors

$$k \int_0^{2\pi} c\lambda \sin^2 \tau \, d\tau = 0. \tag{4.10'}$$

L'équation (4.10′) n'admet pas de solution autre que la solution triviale $c = 0$.

L'équation transcendante (4.10) peut admettre un nombre fini de solutions. Elle peut enfin être une identité valable pour toute valeur de $c$. Cette situation se présentera toutes les fois que la fonction « pertrubatrice » $F$ sera conservative. En effet, supposons que

$F = F(x)$ ; alors

$$J(c) = \int_0^{2\pi} F(c \cos \tau) \sin \tau \, d\tau.$$

Dans ce cas, $F(c \cos \tau)$ est une fonction périodique paire de $\tau$ de période $2\pi$. Elle se décompose donc en une série de Fourier contenant exclusivement des termes en $\cos k\tau$, donc, en vertu de l'orthogonalité de $\sin \tau$ et $\cos k\tau$ on aura pour tout $k$:

$$J(c) \equiv 0.$$

On conviendra dans la suite de considérer que $c$ est une racine non nulle de multiplicité un *) de l'équation (4.10). Dans ce cas, l'équation (4.11) définit une seule valeur du coefficient $g_1(c)$. Donc, l'algorithme de Poincaré nous permet à cette étape de déterminer l'amplitude de la solution génératrice et le premier écart de la fréquence, c'est-à-dire de calculer entièrement l'approximation d'ordre zéro.

Signalons que dans cette modification de la méthode du petit paramètre, la solution de l'équation génératrice ne peut pas être déterminée au pas d'ordre zéro comme dans la méthode « directe » de Poincaré. Cette nuance est essentielle car les solutions de l'équation génératrice ne donnent pas toutes les solutions périodiques: les solutions périodiques ne peuvent exister qu'au voisinage des solutions exclusives de l'équation génératrice. Si l'on se limite au premier terme seulement du développement (4.6), on obtient la solution approchée

$$x \approx x^{(0)}(t) = c \cos \frac{\lambda t}{1 + g_1(c)\varepsilon} \tag{4.12}$$

Mettons la formule (4.12) sous la forme

$$x^{(0)} = c \cos \omega(c) t, \tag{4.13}$$

où

$$\omega(c) = \frac{\lambda}{1 + g_1(c)\varepsilon} \tag{4.14}$$

Cette expression recèle déjà une information très importante: elle permet d'établir la relation existant entre la fréquence et l'amplitude. Les systèmes oscillatoires sont classés en grossiers et moux selon le caractère de cette relation. Si $g_1(c)$ est une fonction monotone strictement décroissante de l'amplitude, alors la fréquence croîtra avec l'amplitude. Le système est dit alors *grossier* (courbe *II* sur la

*) On pourrait envisager des situations plus générales. Mais dans ces conditions il est possible que la solution ne se représente pas par une série (4.6) : la fonction $x(t)$ doit être développée en une série de $\varepsilon$ suivant les puissances fractionnaires.

figure 4.6). Dans le cas contraire le système est dit *mou*. Tous ces résultats ont été établis à $O(\varepsilon)$ près.

Pour obtenir l'approximation suivante, nous devons écrire la solution générale de la deuxième équation du système (4.7). Cette solution contiendra deux nouvelles constantes. L'une d'elles peut être déterminée à partir des conditions initiales, l'autre restera inconnue et peut être trouvée seulement à partir des conditions d'existence d'une solution de l'équation en $x^{(2)}$, et ainsi de suite.

L'analyse de la première approximation nous a conduits à la condition (4.10) d'existence d'une solution, qui est une équation non linéaire. Les conditions d'existence seront des équations linéaires dans les approximations suivantes.

Illustrons la méthode de Poincaré sur deux exemples.

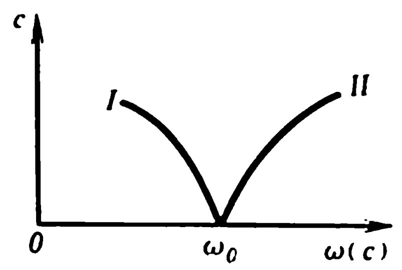



Fig. 4.6

b) *Equation de Duffing.* Nous avons déjà examiné cette équation au paragraphe précédent. Nous avons montré que toutes les solutions de l'équation

$$\ddot{x} + x - x^3 = 0 \tag{4.15}$$

sont périodiques au voisinage de la position d'équilibre $x = 0$. S'agissant de l'équation

$$\ddot{x} + x + x^3 = 0, \tag{4.15'}$$

chacune de ses solutions sera périodique. Nous avons établi ces faits en analysant la structure du plan de phase par des méthodes géométriques. La méthode classique des séries ne nous a pas permis d'obtenir une description satisfaisante de ces solutions. Essayons maintenant d'utiliser la méthode proposée. Mettons tout d'abord les équations (4.15), (4.15′) sous la forme quasi linéaire. Ecrivons à cet effet l'équation

$$\ddot{x} + x = \pm \varepsilon x^3 \tag{4.16}$$

L'équation (4.16) se transforme en (4.15), (4.15′) pour $\varepsilon = 1$. Signalons au passage que ceci est une méthode d'introduction artificielle du paramètre. Le théorème de Poincaré affirme la convergence des séries de $\varepsilon$ uniquement pour $\varepsilon$ assez petit. Mais la « petitesse » de $\varepsilon$ est en fait (nous l'avons vu au paragraphe précédent) définie par une certaine combinaison des paramètres du problème. La possibilité de représenter la solution de l'équation (4.16) sous forme de séries de $\varepsilon$ pour $\varepsilon$ de l'ordre de 1 signifie sans plus que nous choisissons des valeurs initiales assez petites. De ce point de vue, l'introduction artificielle du paramètre dans le système (4.16) équivaut parfaitement

au changement de variables effectué au paragraphe précédent (cf. équation (3.33)).

Le changement (4.3) ramène l'équation (4.16) à la forme

$$\frac{d^2x}{d\tau^2}+x\,[1+2g_1\varepsilon+(2g_2+g_1^2)\,\varepsilon^2+\ \ldots]=$$
$$=\pm\,\varepsilon x^3\,[1+2g_1\varepsilon+(2g_2+g_1^2)\,\varepsilon^2+\ \ldots]. \qquad (4.17)$$

En portant la série (4.6) dans l'équation (4.17), on obtient des équations de la forme (4.7) qui s'écrivent ici

$$\begin{aligned}&\frac{d^2x^{(0)}}{d\tau^2}+x^{(0)}=0,\\ &\frac{d^2x^{(1)}}{d\tau^2}+x^{(1)}=\pm\,x^{(0)3}-2g_1x^{(0)},\\ &\ldots\ldots\ldots\ldots\ldots\ldots\end{aligned} \qquad (4.18)$$

La solution de la première équation du système (4.18) qui vérifie les conditions initiales (4.5) s'écrit

$$x^{(0)}=c\cos\tau.$$

La deuxième équation devient

$$\frac{d^2x^{(1)}}{d\tau^2}+x^{(1)}=\pm\,c^3\cos^3\tau-2g_1\,c\cos\tau. \qquad (4.19)$$

Nous avons deux constantes arbitraires $c$ et $g_1$ qui doivent satisfaire les équations (4.10) et (4.11). Etant donné que dans le cas considéré la « force perturbatrice » $\pm x^3$ est conservative, l'équation (4.10) est identiquement vérifiée par rapport à $c$, c'est-à-dire que quel que soit l'état initial $x^{(0)}=c$, elle engendre une solution périodique: ce fait est corroboré par la conclusion déduite au paragraphe précédent.

Considérons maintenant le problème de la quantité $g_1$. Le mieux est de procéder ainsi. Comme

$$\cos^3\tau=\frac{3}{4}\cos\tau+\frac{1}{4}\cos 3\tau,$$

l'équation (4.19) peut être mise sous la forme suivante:

$$\frac{d^2x^{(1)}}{d\tau^2}+x^{(1)}=c\left(\pm\,\frac{3c^2}{4}-2g_1\right)\cos\tau\pm\frac{c^3}{4}\cos 3\tau. \qquad (4.20)$$

Pour que l'équation (4.20) admette une solution périodique il est nécessaire et suffisant que le développement de son second membre en série de Fourier ne contienne pas cos τ et sin τ. De là il s'ensuit immédiatement que

$$g_1=\pm\,\frac{3}{8}\,c^2. \qquad (4.21)$$

Donc, en faisant $\varepsilon = 1$ on trouve en première approximation la solution périodique sous la forme

$$x = c \cos \omega (c)\, t,$$

où

$$\omega (c) = \frac{1}{1 \pm 3c^2/8}, \tag{4.22}$$

le signe « + » correspondant à l'équation (4.15), le signe « — », à l'équation (4.15'). Donc, si la force de rappel est de la forme $-x + x^3$, le système est mou. Si la force de rappel est égale à $-x - x^3$, le système sera grossier.

Ainsi, dans le dernier cas la fréquence des oscillations croît avec l'amplitude. On peut établir facilement ce caractère des oscillations sans recourir aux calculs. En effet, lorsque l'amplitude $c$ augmente, le rôle du terme non linéaire de l'expression $-x - x^3$ croît, ce qui entraîne un accroissement de la force de rappel et par suite de la fréquence.

En première approximation le mouvement périodique est décrit par une fonction harmonique comme dans la théorie linéaire du pendule. Il existe cependant une distinction fondamentale. La formule (4.22) montre que la fréquence des oscillations dépend de l'amplitude $c$. Dans le cas de l'équation (4.15'), la fréquence croît avec l'amplitude.

c) *Equation de Van der Pol.* Le deuxième exemple illustratif de la théorie des solutions périodiques sera l'équation classique de Van der Pol qui a inspiré une foule de travaux en théorie des oscillations:

$$\ddot{x} + \lambda^2 x = \varepsilon (1 - ax^2)\, \dot{x}.$$

Le changement de variables (4.3) ramène cette équation à la forme

$$\frac{d^2x}{d\tau^2} + x(1 + 2g_1\varepsilon + \ldots) = \varepsilon (1 - ax^2) \frac{dx}{d\tau} \frac{(1 + g_1\varepsilon + \ldots)}{\lambda}. \tag{4.23}$$

En cherchant la solution sous forme de la série

$$x = x^{(0)} + \varepsilon x^{(1)} + \varepsilon^2 x^{(2)} + \ldots$$

on est conduit au système suivant d'équations par rapport à $x^{(0)}$, $x^{(1)}$, . . . :

$$\frac{d^2x^{(0)}}{d\tau^2} + x^{(0)} = 0,$$

$$\frac{d^2x^{(1)}}{d\tau^2} + x^{(1)} = \frac{1}{\lambda} (1 - ax^{(0)2}) \frac{dx^{(0)}}{d\tau} - 2g_1 x^{(0)}, \tag{4.24}$$

. . . . . . . . . . . . . . . . . . . . . . . . .

La première équation du système (4.24), où les fonctions $x^{(i)}$ satisfont les conditions (4.5), nous donne

$$x^{(0)} = c \cos \tau.$$

En portant cette expression dans la deuxième équation, on obtient

$$\frac{d^2x^{(1)}}{d\tau^2}+x^{(1)}=-\frac{1}{\lambda}(1-ac^2\cos^2\tau)\,c\sin\tau-2g_1c\cos\tau.$$

Développons le second membre de cette équation en série de Fourier en remarquant que

$$\cos^2\tau\sin\tau=\frac{1}{4}\sin\tau+\frac{1}{4}\sin 3\tau.$$

On obtient en définitive

$$\frac{d^2x^{(1)}}{d\tau^2}+x^{(1)}=-\frac{c}{\lambda}\left\{\sin\tau\left(1-\frac{ac^2}{4}\right)-\frac{ac^2}{4}\sin 3\tau\right\}-2g_1c\cos\tau. \tag{4.25}$$

Pour que l'équation (4.25) admette des solutions périodiques de période $2\pi$, il est nécessaire et suffisant que le développement du second membre en série de Fourier ne contienne pas les premiers harmoniques, c'est-à-dire que les coefficients de $\sin\tau$ et de $\cos\tau$ soient nuls. Ceci nous fournit deux équations pour la détermination de $c$ et $g_1$, qui nous donnent

$$2g_1=0,\qquad c\,(1-ac^2/4)=0.$$

Donc, $g_1=0$. La deuxième équation admet trois racines:

$$c_1=0,\quad c_2=2/\sqrt{a},\quad c_3=-2/\sqrt{a}. \tag{4.26}$$

Donc, l'équation (4.25) admet deux régimes stationnaires. L'un d'eux est l'état d'équilibre $x\equiv 0$, l'autre, un régime auto-oscillatoire. Si l'on se limite à la première approximation, ce mouvement sera décrit par la formule

$$x=\pm\frac{2}{\sqrt{a}}\cos\tau.$$

La méthode de Poincaré de recherche des solutions périodiques est loin d'être universelle. Tout d'abord, elle permet d'étudier les processus oscillatoires seulement dans les systèmes quasi linéaires, c'est-à-dire pour de petites valeurs de $\varepsilon$. Ainsi, même l'équation relativement simple de Van der Pol possède de nombreuses propriétés qui ne peuvent être étudiées par la théorie développée (par exemple, l'existence éventuelle de solutions périodiques proches des oscillations discontinues).

Par ailleurs, la méthode de Poincaré vise simplement la recherche de solutions périodiques et ne permet pas d'étudier les processus d'établissement des mouvements, la stabilité, etc.

Enfin, cette méthode est bien élaborée surtout pour les systèmes du second ordre ou pour les systèmes de structure « normale », c est-à-

dire de la forme

$$\ddot{x}_i + \omega_i^2 x_i = \varepsilon f_i (x_1, \ldots, x_n, \dot{x}_1, \ldots, \dot{x}_n), \quad i = 1, \ldots, n. \tag{4.27}$$

Mais la réduction d'un système arbitraire, même quasi linéaire, de la forme

$$\dot{z} + Az = \varepsilon F(z)$$

à la forme (4.27) est un problème dont l'analyse numérique est assez compliquée.

## § 5. Méthode de moyennisation

Dans ce chapitre nous avons déjà rencontré des systèmes d'équations contenant un petit paramètre auxquels l'application directe des méthodes de développement de la solution en une série du petit paramètre (développement qui est théoriquement possible en vertu du théorème de Poincaré) n'a fourni aucun résultat utile pour la pratique. L'un de ces systèmes était de la forme *)

$$\begin{aligned} \dot{x} &= \varepsilon X(x, y, \varepsilon), \\ \dot{y} &= \omega(x) + \varepsilon Y(x, y, \varepsilon), \end{aligned} \tag{5.1}$$

où $x$ est un vecteur, $y$, un scalaire, $X$ et $Y$, des fonctions périodiques de période $2\pi$. Au § 3 nous avons montré comment les systèmes de forme (5.1) étaient reliés aux systèmes contenant des éléments oscillatoires. Les difficultés soulevées par l'intégration numérique des systèmes (5.1) sont dues au fait que ces systèmes renferment des variables « lentes » $x$ et une variable « rapide » $y$ (c'est-à-dire une variable dont la dérivée est de l'ordre de $O(1)$ et non pas de $O(\varepsilon)$ comme celle de $x$) et le pas d'intégration est imposé par la variable rapide. Pour cette raison, le système (5.1) doit être intégré avec un petit pas, ce qui apporte des complications lorsque l'intervalle de temps est élevé: longue occupation de l'ordinateur, perte de précision, etc. L'application de la méthode de Poincaré, comme nous l'avons déjà vu, n'oblitère pas ces obstacles.

REMARQUE. A signaler que le problème de l'affinement du pas d'intégration numérique est presque la plus grosse pierre d'achoppement pour le programmeur. R. Bellman a été même jusqu'à parler de la « malédiction de la dimension ». La présence des termes oscillatoires n'est pas non plus pour arranger des choses.

L'idée de la séparation des mouvements lents et des mouvements rapides avancée au début des années trente déjà par N. Krylov et N. Bogolioubov peut être très fructueuse dans la simplification de

*) De tels systèmes sont dits *systèmes à phase tournante*.

l'analyse des systèmes de la forme (5.1). L'évolution des variables lentes, par exemple les variations des paramètres du milieu dans l'exemple du § 3, présente en principe un grand intérêt pour l'analyste. Si l'on réussissait à mettre ces variables en évidence, on obtiendrait une méthode très économique d'analyse des plus importantes caractéristiques du système.

Mais ceci est impossible à faire directement. Il est impossible, par exemple, de fixer tout simplement la valeur de la variable rapide dans la première équation du système (5.1), c'est-à-dire de négliger la contribution de l'intégrale des composantes variant rapidement du système. Signalons que l'interdépendance des variations des deux variables fait que non seulement $y(t)$, mais $x(t)$ aussi revêt un caractère oscillatoire.

Ceci nous suggère un changement de variables susceptible de séparer les mouvements rapides des mouvements lents. Mais dans le cas général il est impossible de trouver un tel changement. Force est donc de chercher une substitution qui donne une solution approchée, si possible asymptotique, de ce problème. Ces changements sont cherchés sous la forme des séries

$$\begin{aligned} x &= \bar{x} + \varepsilon u_1(\bar{x}, \bar{y}) + \varepsilon^2 u_2(\bar{x}, \bar{y}) + \ldots, \\ y &= \bar{y} + \varepsilon v_1(\bar{x}, \bar{y}) + \varepsilon^2 v_2(\bar{x}, \bar{y}) + \ldots \end{aligned} \tag{5.2}$$

Diverses procédures ont été mises au point pour composer des équations pour $\bar{x}$ et $\bar{y}$ ne contenant plus les termes oscillatoires et pour déterminer les termes supplémentaires $u_i$ et $v_i$. La théorie des transformations de la forme (5.2) est aujourd'hui un chapitre bien élaboré des mathématiques appliquées. Il s'avère en particulier que dans le cas général les séries (5.2) divergent, mais elles sont asymptotiques sous certaines conditions. La dernière assertion signifie que des tranches finies de ces séries peuvent servir à approcher la solution exacte sur un grand intervalle de temps de l'ordre de $1/\varepsilon$, et l'erreur de calcul sur un tel intervalle sera de l'ordre du dernier terme éliminé.

Nous glisserons sur la discussion de ces questions assez profondes de théorie, nous bornant seulement à la technique de réalisation de ces transformations. Considérons un schéma d'approximations successives (cf. [7]) proposé par l'auteur. Ce schéma est plus économique sur le plan des calculs que l'utilisation de séries (5.2) et n'implique pas les conditions contraignantes de continuité et de dérivabilité des fonctions figurant dans les équations.

a) *Schéma d'approximations successives.* Cherchons la transformation des variables sous la forme

$$\begin{aligned} x &= \bar{x} + \varepsilon u(\bar{x}, \bar{y}, \varepsilon), \\ y &= \bar{y} + \varepsilon v(\bar{x}, \bar{y}, \varepsilon). \end{aligned} \tag{5.3}$$

Exigeons que la fonction vectorielle $\bar{x}(t)$ vérifie un système d'équations ne contenant pas la variable « rapide » $\bar{y}(t)$. Exigeons de même que la fonction scalaire $\bar{y}(t)$ soit solution d'une équation ne contenant pas $\bar{y}(t)$ au second membre. En d'autres termes, exigeons que les nouvelles variables $\bar{x}(t)$ et $\bar{y}(t)$ soient solutions d'équations de la forme:

$$\begin{aligned} \dot{\bar{x}} &= \varepsilon A(\bar{x}, \varepsilon), \\ \dot{\bar{y}} &= \omega(\bar{x}) + \varepsilon B(\bar{x}, \varepsilon). \end{aligned} \tag{5.4}$$

Les fonctions $A$ et $B$ ne sont pas connues *a priori*.

Si l'on réussit à trouver la transformation désirée, on est conduit au système d'équations (5.4) qui est bien plus simple que l'initial. En effet, dans le système (5.4) le mouvement lent dont la vitesse est de l'ordre de $O(\varepsilon)$ est entièrement séparé du mouvement rapide dont la vitesse est de l'ordre de $O(1)$. Donc, le système d'équations définissant le vecteur $\bar{x}$ s'intègre indépendamment de l'équation pour $\bar{y}$. Le pas d'intégration par rapport à $t$ peut être pris grand, car la dérivée $\dot{\bar{x}}$ est petite et le second membre de l'équation pour $\bar{x}$ ne dépend pas de $\bar{y}$. Une fois $\bar{x}$ connu, on peut déterminer $\bar{y}$ par une intégration.

Cette opération est assez économique, car on aura à intégrer une fonction variant lentement $\omega(\bar{x}) + \varepsilon B\bar{x}(\varepsilon)$.

Observons que les fonctions $\bar{x}$ et $\bar{y}$, solutions des équations « séparées », sont bien les nouvelles variables: non seulement elles ne sont pas identiques à $x$ et $y$, mais leur connaissance ne suffit encore pas à déterminer les $x$ et $y$ initiales. En même temps elles diffèrent d'elles, comme le montrent les formules (5.3), d'une quantité de l'ordre de $O(\varepsilon)$. Donc, toutes les interactions des processus $x(t)$ et $y(t)$ sont du même ordre. Si l'on réussit à construire la transformation (5.3), on peut, en modifiant le système initial d'une quantité de l'ordre de $O(\varepsilon)$, lever le principal obstacle à son analyse numérique.

Ces remarques faites, passons à la description de l'algorithme sans oublier qu'il nous faut encore trouver des méthodes de calcul des fonctions $A$, $B$, $u$ et $v$.

Une fois qu'on a trouvé $\bar{x}(t)$ et $\bar{y}(t)$, on doit déterminer les fonctions $u(\bar{x}, \bar{y}, \varepsilon)$ et $v(\bar{x}, \bar{y}, \varepsilon)$. La détermination de ces fonctions risque de donner lieu à un grand arbitraire dont on peut se dédouaner par l'introduction de contraintes supplémentaires sur ces fonctions. Choisissons-les parmi les fonctions bornées lorsque $\bar{y}(t) \to \infty$. Cette contrainte est naturelle, car elle permet de traiter $\varepsilon u$ et $\varepsilon v$ comme des quantités de l'ordre de $O(\varepsilon)$ pour tout $\bar{y}$.

Ainsi, la recherche de la transformation (5.3) se ramène à la détermination de fonctions $u(\bar{x}, \bar{y}, \varepsilon)$, $v(\bar{x}, \bar{y}, \varepsilon)$, $A(\bar{x}, \varepsilon)$ et $B(\bar{x}, \varepsilon)$. Pour déterminer ces quantités, portons les expressions (5.3), (5.4) dans les équations (5.1). En simplifiant par $\varepsilon$, on obtient

$$\begin{gathered} A(\bar{x}, \varepsilon)+\varepsilon \frac{\partial u}{\partial \bar{x}} A(\bar{x}, \varepsilon)+\frac{\partial u}{\partial \bar{y}}(\omega(\bar{x})+\varepsilon B(\bar{x}, \varepsilon))= \\ =X(\bar{x}+\varepsilon u, \bar{y}+\varepsilon v, \varepsilon), \\ \omega(\bar{x})+\varepsilon B(\bar{x}, \varepsilon)+\varepsilon^2 \frac{\partial v}{\partial \bar{x}} A(\bar{x}, \varepsilon)+\varepsilon \frac{\partial v}{\partial \bar{y}}(\omega(\bar{x})+\varepsilon B(\bar{x}, \varepsilon))= \\ =\omega(\bar{x}+\varepsilon u)+\varepsilon Y(\bar{x}+\varepsilon u, \bar{y}+\varepsilon v, \varepsilon). \end{gathered} \tag{5.5}$$

Traitons le système (5.5) par la méthode des approximations successives. Mettons-le à cet effet sous la forme

$$\begin{gathered} \frac{\partial u}{\partial \bar{y}} \omega(\bar{x})=g(\bar{x}, \bar{y}, u, v, A, B)-A(\bar{x}, \varepsilon), \\ \frac{\partial v}{\partial \bar{y}} \omega(\bar{x})=h(\bar{x}, \bar{y}, u, v, A, B)-B(\bar{x}, \varepsilon), \end{gathered} \tag{5.6}$$

où

$$g(\bar{x}, \bar{y}, u, v, A, B)=X(\bar{x}+\varepsilon u, \bar{y}+\varepsilon v, \varepsilon)-\varepsilon \frac{\partial u}{\partial \bar{x}} A-\varepsilon \frac{\partial u}{\partial \bar{y}} B,$$

$$h=\frac{\omega(\bar{x}+\varepsilon u)-\omega(\bar{x})}{\varepsilon}+Y(\bar{x}+\varepsilon u, \bar{y}+\varepsilon v, \varepsilon)-\varepsilon \frac{\partial v}{\partial \bar{x}} A-\varepsilon \frac{\partial v}{\partial \bar{y}} B.$$

Utilisons le schéma itératif suivant pour résoudre (5.6):

$$\begin{gathered} \frac{\partial u^{(k)}}{\partial \bar{y}} \omega(\bar{x})=g_k-A^{(k)}, \\ \frac{\partial v^{(k)}}{\partial \bar{y}} \omega(\bar{x})=h_k-B^{(k)}, \\ g_k=g(\bar{x}, \bar{y}, u^{(k-1)}, v^{(k-1)}, A^{(k-1)}, B^{(k-1)}), \\ h_k=h(\bar{x}, \bar{y}, u^{(k-1)}, v^{(k-1)}, A^{(k-1)}, B^{(k-1)}). \end{gathered} \tag{5.7}$$

On admettra que

$$u^{(0)}=v^{(0)}=A^{(0)}=B^{(0)}=0.$$

Donc, en vertu des formules (5.7) les équations de la première approximation s'écrivent

$$\begin{gathered} \frac{\partial u^{(1)}}{\partial \bar{y}} \omega(\bar{x})=X(\bar{x}, \bar{y}, \varepsilon)-A^{(1)}(\bar{x}, \varepsilon), \\ \frac{\partial v^{(1)}}{\partial \bar{y}} \omega(\bar{x})=Y(\bar{x}, \bar{y}, \varepsilon)-B^{(1)}(\bar{x}, \varepsilon). \end{gathered} \tag{5.8}$$

Ainsi, la recherche de la transformation se ramène à la résolution successive de systèmes d'équations aux dérivées partielles (5.7). Discutons maintenant ce processus sur l'exemple du système (5.8). Tous les autres systèmes d'équations déduits de (5.7) présentent la même structure et leur étude peut être conduite de la même manière.

Considérons la première équation du système (5.8). C'est une équation aux dérivées partielles du premier ordre, résolue par rapport à la dérivée de la fonction inconnue $u^{(1)}(\bar{x}, \bar{y}, \varepsilon)$ et ne contenant pas la dérivée $\dot{\bar{y}}$ de l'autre argument. Cette équation est encore caractérisée par le fait que son second membre contient une fonction inconnue $A^{(1)}(\bar{x}, \varepsilon)$ qui reste à déterminer.

Une intégration de la première équation du système (5.8) par rapport à $\bar{y}$ nous donne

$$u^{(1)}(\bar{x}, \bar{y}, \varepsilon) = \frac{1}{\omega(\bar{x})} \int_{\bar{y}_0}^{\bar{y}} \{X(\bar{x}, \bar{y}, \varepsilon) - A^{(1)}(\bar{x}, \varepsilon)\} d\bar{y} + \varphi(\bar{x}), \tag{5.9}$$

où $\varphi(\bar{x})$ est la constante d'intégration, c'est-à-dire une fonction arbitraire de $\bar{x}$. Pour déterminer la fonction inconnue $A^{(1)}(\bar{x}, \varepsilon)$ il faut introduire une condition supplémentaire. Cette condition a déjà été énoncée. Qu'on se souvienne en effet qu'on a convenu de chercher la solution dans la classe des fonctions bornées:

$$\lim_{\bar{y} \to \infty} |u^{(i)}(\bar{x}, \bar{y}, \varepsilon)| < \infty$$

pour tout $\bar{x}$. L'intégrant de (5.9) est une fonction périodique de $\bar{y}$, puisque par hypothèse $X$ est une fonction périodique de $\bar{y}$ et $A^{(1)}(\bar{x}, \varepsilon)$ ne dépend pas de $\bar{y}$. Cette fonction est $2\pi$-périodique en $\bar{y}$. Supposons maintenant que sa valeur moyenne sur une période n'est pas nulle, c'est-à-dire que

$$\overline{X - A^{(1)}} = \frac{1}{2\pi} \int_{\bar{y}_0}^{\bar{y}_0 + 2\pi} [X(\bar{x}, \bar{y}, \varepsilon) - A^{(1)}(\bar{x}, \varepsilon)] d\bar{y} = c(\bar{x}) \neq 0. \tag{5.10}$$

Il est alors évident que

$$\lim_{k \to \infty} u^{(1)}(\bar{x}, \bar{y}_0 + 2\pi k, \varepsilon) = \frac{1}{\omega(\bar{x})} \lim_{k \to \infty} 2\pi k c(\bar{x}) = \pm\infty.$$

Le signe dépend de celui de la fonction $c(\bar{x})$. Donc, pour que la fonction $u^{(1)}$ soit bornée, il est nécessaire et suffisant que

$$c(\bar{x}) = 0 \tag{5.11}$$

pour tout $\bar{x}$.

L'intégrale de l'équation (5.9) contient une seule fonction inconnue $A^{(1)}(\bar{x}, \varepsilon)$. La condition (5.11) suffit pour la déterminer. Comme

$$c(\bar{x}) = \frac{1}{2\pi} \int_{\bar{y}_0}^{\bar{y}_0+2\pi} X(\bar{x}, \bar{y}, \varepsilon)\, d\bar{y} - A^{(1)}(\bar{x}, \varepsilon),$$

la condition (5.11) donne

$$A^{(1)}(\bar{x}, \varepsilon) = \bar{X}(\bar{x}, \varepsilon), \tag{5.12}$$

où

$$\bar{X} = \frac{1}{2\pi} \int_{\bar{y}}^{\bar{y}_0+2\pi} X(\bar{x}, \bar{y}, \varepsilon)\, d\bar{y}.$$

La deuxième équation du système (5.8) est exactement analogue à la première. En reprenant les calculs, on obtient finalement les quadratures suivantes qui représentent les fonctions cherchées $u^{(1)}(\bar{x}, \bar{y}, \varepsilon)$ et $v^{(1)}(\bar{x}, \bar{y}, \varepsilon)$ :

$$\begin{aligned} u^{(1)}(\bar{x}, \bar{y}, \varepsilon) &= \frac{1}{\omega(\bar{x})}\left[\int_{\bar{y}_0}^{\bar{y}} X(\bar{x}, \bar{y}, \varepsilon)\, d\bar{y} - \bar{X}\bar{y}\right] + \varphi_1(\bar{x}), \\ v^{(1)}(\bar{x}, \bar{y}, \varepsilon) &= \frac{1}{\omega(\bar{x})}\left[\int_{\bar{y}_0}^{\bar{y}} Y(\bar{x}, \bar{y}, \varepsilon)\, d\bar{y} - \bar{Y}\bar{y}\right] + \psi_1(\bar{x}). \end{aligned} \tag{5.13}$$

Les expressions (5.13) contiennent les fonctions $\varphi_1(\bar{x})$ et $\psi_1(\bar{x})$ qui sont les « constantes » d'intégration. On pourrait proposer plusieurs méthodes différentes pour choisir les fonctions arbitraires $\varphi_1(\bar{x})$ et $\psi_1(\bar{x})$. Il s'avère toutefois que la précision de la solution approchée ne dépend pas du choix de ces fonctions *). Donc, la transformation (5.3) ne conduit pas à un résultat univoque.

REMARQUE. La non-unicité des approximations asymptotiques est caractéristique de toutes les théories asymptotiques. Cette non-unicité ne signifie qu'une chose : il existe toute une famille de fonctions approchant la solution avec la même précision en ce sens que l'écart entre la solution exacte et la solution approchée tend vers 0 avec $\varepsilon$ indépendamment de la fonction utilisée pour approcher la solution.

Pour condition à la détermination des fonctions arbitraires $\varphi_1(\bar{x})$ et $\psi_1(\bar{x})$, on peut exiger que le système initial (5.1) et le système séparé (5.4) satisfassent les mêmes conditions initiales. De cette condition et des formules (5.3), on déduit immédiatement que

*) Pour plus de détails cf. [7].

$u^{(1)}(\bar{x}_0, \bar{y}_0, \varepsilon) = 0, v^{(1)}(\bar{x}_0, \bar{y}_0, \varepsilon) = 0$ d'où il s'ensuit que

$$\varphi_1(\bar{x}_0) = 0, \quad \psi_1(\bar{x}_0) = 0. \tag{5.14}$$

Si par ailleurs $x = \bar{x}$ pour $y = \bar{y}_0$, alors $\varphi_1(\bar{x}) = 0$, $\psi_1(\bar{x}) = 0$. Dans la suite nous nous servirons principalement des formules (5.14). Les autres équations du système (5.7) s'étudient de façon analogue. Les égalités (5.14) aidant, on obtient les formules suivantes pour les solutions de ces équations:

$$\begin{gathered} A^{(k)}(\bar{x}, \varepsilon) = \bar{g}_k(\bar{x}, \varepsilon) = \frac{1}{2\pi} \int_{\bar{y}_0}^{\bar{y}_0+2\pi} g_k(\bar{x}, \bar{y}, \varepsilon)\, d\bar{y}, \\ B^{(k)}(\bar{x}, \varepsilon) = \bar{h}_k(\bar{x}, \varepsilon) = \frac{1}{2\pi} \int_{\bar{y}_0}^{\bar{y}_0+2\pi} h_k(\bar{x}, \bar{y}, \varepsilon)\, d\bar{y}, \\ u^{(k)}(\bar{x}, \bar{y}, \varepsilon) = \frac{1}{\omega(\bar{x})} \left\{ \int_{\bar{y}_0}^{\bar{y}} g_k(\bar{x}, \bar{y}, \varepsilon)\, d\bar{y} - \bar{g}_k(\bar{x}, \varepsilon)\, \bar{y} \right\}, \\ v^{(k)}(\bar{x}, \bar{y}, \varepsilon) = \frac{1}{\omega(\bar{x})} \left\{ \int_{\bar{y}_0}^{\bar{y}} h_k(\bar{x}, \bar{y}, \varepsilon)\, d\bar{y} - \bar{h}_k(\bar{x}, \varepsilon)\, \bar{y} \right\}. \end{gathered} \tag{5.15}$$

Observons que les formules (5.15) renferment des intégrales des fonctions variant rapidement $\bar{y}(t)$, mais que l'intégration a lieu sur une période seulement.

b) *Construction de la solution approchée.* Le processus itératif proposé est généralement divergent. Mais avec un nombre fini d'itérations on peut composer des expressions qui approchent avec une certaine précision la solution du problème initial.

Les théorèmes prouvés par N. Bogolioubov et son école montrent que les agrégats obtenus par un nombre fini d'itérations fournissent une approximation de l'ordre de $1/\varepsilon$ sur de grands intervalles de temps. Discutons maintenant la structure de la solution approchée acquise à l'aide des formules (5.15).

Supposons donc qu'on ait calculé $A^{(n)}$ et $B^{(n)}$. Désignons par $\bar{x}_n$ une solution de l'équation différentielle

$$\dot{\bar{x}}_n = \varepsilon A^{(n)}(\bar{x}_n, \varepsilon). \tag{5.16}$$

Calculons maintenant $\bar{y}_n$ par une intégration:

$$\bar{y}_n = y(0) + \int_0^t \{\omega(\bar{x}_n) + \varepsilon B^{(n)}(\bar{x}_n, \varepsilon)\}\, dt. \tag{5.17}$$

17—01523

Portons les fonctions $\bar{x}_n$ et $\bar{y}_n$ dans les expressions (5.15) et définissons les fonctions $u^{(k)}$ et $v^{(k)}$. Prenons les fonctions suivantes

$$\begin{aligned} x(t) &= \bar{x}_n(t) + \varepsilon u^{(N_1)}(\bar{x}_n, \bar{y}_n, \varepsilon), \\ y(t) &= \bar{y}_n(t) + \varepsilon v^{(N_2)}(\bar{x}_n, \bar{y}_n, \varepsilon) \end{aligned} \tag{5.18}$$

pour solution approchée du problème initial. Le choix des nombres $N_1$ et $N_2$ doit obéir à des règles bien précises. Il faut en effet que les formules approchées (5.18) ne contiennent pas de termes qui puissent être négligés sans perte de précision.

Considérons maintenant une approximation élémentaire. En se bornant aux premiers termes, on peut écrire

$$x(t) = \bar{x}_1, \quad y(t) = \bar{y}_1,$$

et de plus

$$\dot{\bar{x}}_1 = \varepsilon \bar{X}(\bar{x}_1, \varepsilon), \quad \dot{\bar{y}}_1 = \omega(\bar{x}_1) + \varepsilon \bar{Y}(\bar{x}_1, \varepsilon). \tag{5.19}$$

Penchons-nous à présent sur la précision de l'approximation définie par les formules (5.19). En calculant $A^{(1)}$ on s'est servi d'une équation dans laquelle on a négligé quelques termes d'ordre $O(\varepsilon)$: on n'a conservé que les termes qui ne contenaient pas le facteur $\varepsilon$. On a donc commis une erreur de l'ordre de $O(\varepsilon)$ dans le calcul de la dérivée, c'est-à-dire qu'on peut écrire

$$\dot{\bar{x}} = \varepsilon\,(\bar{X}(\bar{x}, \varepsilon) + O(\varepsilon)). \tag{5.20}$$

On se propose d'utiliser l'approximation (5.19) de la solution exacte sur un grand intervalle de temps, de l'ordre de $O(1/\varepsilon)$. Donc, en calculant $\bar{x}_n$ à l'aide de la formule (5.16) on commet une erreur du même ordre que celle qui affecte la dérivée, c'est-à-dire que

$$\bar{x} = \bar{x}_1 + O(\varepsilon). \tag{5.21}$$

REMARQUE. Signalons que sans nuire à la précision on peut remplacer l'équation (5.20) par l'équation

$$\dot{\bar{x}} = \varepsilon\,(\bar{X}(\bar{x}, 0) + O(\varepsilon)).$$

Voyons maintenant la deuxième équation (5.19). Supposons que l'erreur sur le vecteur $\bar{x}$ est de l'ordre de $\varepsilon$; la quantité $\omega(\bar{x})$ sera alors déterminée avec la même erreur:

$$\omega(\bar{x}) = \omega(\bar{x}_1) + O(\varepsilon).$$

Donc, l'intégration de $\omega(\bar{x})$ sur un intervalle de longueur de l'ordre de $O(1/\varepsilon)$ donne lieu à l'estimation

$$\int_0^t \omega(\bar{x})\,dt = \int_0^t \omega(\bar{x}_1)\,dt + O(1).$$

Donc, l'intégrant de la formule (5.17) ne doit renfermer que les termes dont l'ordre est inférieur à celui de l'erreur affectant le calcul de $\bar{x}_n$. Dans le cas où $n = 1$ seul le premier terme doit être conservé. De ce point de vue, la deuxième équation (5.19) est écrite avec une précision « redondante ». Il faut la remplacer par l'équation

$$\dot{\bar{y}}_1 = \omega(\bar{x}_1). \tag{5.19'}$$

On estime de façon analogue les approximations d'ordre plus élevé. Ainsi

$$\begin{gathered} \bar{x} = \bar{x}_2 + O(\varepsilon^2), \\ \dot{\bar{x}}_2 = \varepsilon A^{(2)}, \quad \dot{\bar{y}}_2 = \omega(\bar{x}_2) + \varepsilon \bar{Y}. \end{gathered} \tag{5.22}$$

Donc, le calcul des mouvements lents est plus précis (d'un ordre) à un même niveau d'itération que celui de la variable rapide *). Dans le cas où $\omega$ ne dépend pas de $x$, les quantités $\bar{x}$ et $\bar{y}$ sont déterminées avec la même précision et la deuxième formule (5.19) est valable; quant à la dernière équation (5.22), il faut la remplacer par l'équation

$$\dot{\bar{y}}_2 = \omega(\bar{x}_2) + \varepsilon B^{(2)}.$$

Il faut procéder exactement de la même façon pour le calcul des fonctions $u$ et $v$. On obtient en définitive les formules suivantes des transformations. Pour la première approximation, on a

$$x = \bar{x}_1, \qquad y = \bar{y}_1.$$

La deuxième approximation est donnée par les formules

$$\begin{gathered} x = \bar{x}_2 + \varepsilon u^{(1)}(\bar{x}_1, \bar{y}_1, \varepsilon), \\ y = \begin{cases} \bar{y}_2, & \text{si } \omega = \omega(x), \\ \bar{y}_2 + \varepsilon v^{(1)}(\bar{x}_1, \bar{y}_1, \varepsilon), & \text{si } \omega = \text{const.} \end{cases} \end{gathered} \tag{5.23}$$

Pour calculer la deuxième approximation, il faut effectuer les opérations suivantes.

1) Résoudre le système pour la première approximation:

$$\dot{\bar{x}}_1 = \bar{X}(\bar{x}_1, \varepsilon), \qquad \dot{\bar{y}}_1 = \omega(\bar{x}_1).$$

*) On a déjà dit que l'étude des mouvements lents présentait le plus grand intérêt dans les problèmes pratiques C'est pourquoi le fait qu'au niveau d'une même approximation les mouvements lents sont calculés avec une meilleure précision n'apporte généralement pas de complication.

2) Calculer $u^{(1)}$ et $v^{(1)}$ avec les formules (5.13) :

$$u^{(1)}=\frac{1}{\omega(\bar{x}_1)}\int_{\bar{y}_0}^{\bar{y}}\{X(\bar{x}_1,\ \bar{y}_1,\ \varepsilon)-\overline{X}(\bar{x}_1,\ \varepsilon)\}\,d\bar{y}_1.$$

$$v^{(1)}=\frac{1}{\omega(\bar{x}_1)}\int_{\bar{y}_0}^{\bar{y}}\{Y(\bar{x}_1,\ \bar{y}_1,\ \varepsilon)-\overline{Y}(\bar{x}_1,\ \varepsilon)\}\,d\bar{y}_1.$$

3) Composer les expressions de $g_2$ et $h_2$ :

$$g_2(\bar{x}_1,\ \bar{y}_1,\ \varepsilon)=X(\bar{x}_1+\varepsilon u^{(1)},\ \bar{y}_1+\varepsilon v^{(1)},\ \varepsilon)-\\ -\varepsilon\frac{\partial u^{(1)}}{\partial\bar{x}_1}\overline{X}(\bar{x}_1,\ \varepsilon)-\varepsilon\frac{\partial u^{(1)}}{\partial\bar{y}_1}\overline{Y}(\bar{x}_1,\ \varepsilon),$$

$$h_2(\bar{x}_1,\ \bar{y}_1,\ \varepsilon)=Y(\bar{x}_1+\varepsilon u^{(1)},\ \bar{y}_1+\varepsilon v^{(1)},\ \varepsilon)+\frac{\omega(\bar{x}_1+\varepsilon u^{(1)})-\omega(\bar{x}_1)}{\varepsilon}-\\ -\varepsilon\frac{\partial v^{(1)}}{\partial\bar{x}_1}\overline{X}(\bar{x}_1,\ \varepsilon)-\varepsilon\frac{\partial v^{(1)}}{\partial\bar{y}_1}\overline{Y}(\bar{x}_1,\ \varepsilon).$$

4) Calculer $A^{(2)}$ et $B^{(2)}$ :

$$A^{(2)}(\bar{x}_1,\ \varepsilon)=\overline{g_2(\bar{x}_1,\ \bar{y}_1,\ \varepsilon)}=\frac{1}{2\pi}\int_{\bar{y}_0}^{\bar{y}_0+2\pi}g_2(\bar{x}_1,\ \bar{y}_1,\ \varepsilon)\,d\bar{y}_1,$$

$$B^{(2)}(\bar{x}_1,\ \varepsilon)=\overline{h_2(\bar{x}_1,\ \bar{y}_1,\ \varepsilon)}=\frac{1}{2\pi}\int_{\bar{y}_0}^{\bar{y}_0+2\pi}h_2(\bar{x}_1,\ \bar{y}_1,\ \varepsilon)\,d\bar{y}_1.$$

5) Intégrer le système

$$\dot{\bar{x}}_2=\varepsilon A^{(2)},$$

$$\dot{\bar{y}}_2=\begin{cases}\omega(\bar{x}_2)+\varepsilon B^{(1)}, & \text{si } \omega=\omega(x),\\ \omega+\varepsilon B^{(2)}, & \text{si } \omega=\text{const.}\end{cases}$$

On constate que le volume des calculs croît brusquement avec l'ordre de l'approximation. Aussi dans la plupart des cas est-on contraint de limiter l'analyse aux seules équations pour la première approximation. Il existe toutefois des problèmes techniques importants dans lesquels les plus remarquables propriétés ne sont mises en lumière que par une étude des approximations d'ordre plus élevé *).

Récapitulons. Nous avons étudié une classe de systèmes de la forme (5.1) dont les seconds membres étaient des fonctions périodiques de la variable rapide $y$. Nous avons déjà signalé que cette classe était

*) Exemple, les mouvements giratoires d'un solide dans un champ de gravitation. Pour plus de détails voir [7].

assez vaste et se rencontrait souvent dans les applications. Font partie de cette classe tous les systèmes de la forme

$$\dot{x} = \varepsilon X(x, t, \varepsilon) \tag{5.24}$$

dont les seconds membres sont des fonctions périodiques du temps $t$. Le système d'équations (5.24) est un cas particulier du système (5.1). Pour s'en assurer il suffit d'introduire la variable rapide $y$:

$$dy/dt = 1. \tag{5.25}$$

Appliquons l'algorithme développé pour résoudre le système (5.24), (5.25). Si les impératifs de la précision nous permettent de nous limiter à la première approximation, alors la résolution de ce système se ramène à celle du système moyennisé

$$\dot{\bar{x}} = \varepsilon \bar{X}(\bar{x}, \varepsilon), \tag{5.26}$$

où

$$\bar{X} = \frac{1}{T} \int_0^T X(\bar{x}, y, \varepsilon)\, dy,$$

et $T$ est la période de $X$ par rapport à $y$. L'écart entre la solution exacte et la solution approchée sera de l'ordre de $O(\varepsilon)$ pour un intervalle d'intégration de longueur $1/\varepsilon$. Lorsque $\varepsilon$ diminue, la précision croît, de même d'ailleurs que l'intervalle d'intégration.

Mais la théorie exposée trouve sa plus importante application dans les systèmes contenant des termes oscillatoires. Au § 3, nous avons vu comment ces systèmes se ramenaient à la forme (5.1). Nous reviendrons sur les systèmes à termes oscillatoires à la fin de ce paragraphe. L'application de la technique de moyennisation n'a pas pour seul effet de rabaisser l'ordre du système, le plus important c'est qu'elle ramène son étude à l'analyse d'un autre système faisant intervenir uniquement des variables lentes. Donc, l'intégration numérique de tels systèmes ne soulève pas de grosses difficultés et peut être effectuée d'une façon économique.

La méthode de moyennisation se prête à de nombreuses généralisations importantes. Dans les systèmes (5.1) on peut notamment renoncer aux conditions de périodicité des seconds membres par rapport à la variable rapide. Si par exemple les fonctions $X$ et $Y$ sont quasi périodiques, alors le système d'approximation sera le suivant:

$$\begin{aligned} \dot{x} &= \varepsilon \lim_{\alpha\to\infty} \frac{1}{2\alpha} \int_{-\alpha}^{+\alpha} X(x, y, \varepsilon)\, dy, \\ \dot{y} &= \omega + \varepsilon \lim_{\alpha\to\infty} \frac{1}{2\alpha} \int_{-\alpha}^{+\alpha} Y(x, y, \varepsilon)\, dy. \end{aligned} \tag{5.27}$$

c) *Etude de la stabilité par la méthode de moyennisation.* Les raisonnements développés fournissent une méthode souple et très efficace d'analyse des systèmes. Ils permettent en particulier de trouver les solutions périodiques de nombreux problèmes examinés dans le paragraphe précédent. Citons un exemple qui met en lumière ces possibilités. Considérons l'équation de Van der Pol (cf. § 4):

$$\ddot{x} + \lambda^2 x = \varepsilon (1 - ax^2)\, \dot{x}, \tag{5.28}$$

et faisons le changement de variables

$$x = c \cos \varphi, \quad \dot{x} = -c\lambda \sin \varphi. \tag{5.29}$$

En dérivant la première relation et en l'égalant à la seconde, on obtient

$$\dot{c} \cos \varphi - c\dot{\varphi} \sin \varphi + \lambda c \sin \varphi = 0. \tag{5.30}$$

En dérivant la deuxième relation (5.29) et en portant dans (5.28), on trouve

$$-\dot{c}\lambda \sin \varphi - \lambda c\dot{\varphi} \cos \varphi + c\lambda^2 \cos \varphi = \\ = -\varepsilon c (1 - ac^2 \cos^2 \varphi)\, \lambda \sin \varphi. \tag{5.31}$$

La résolution du système (5.30), (5.31) par rapport à $\dot{c}$ et $\dot{\varphi}$ donne

$$\begin{aligned} \dot{c} &= \varepsilon c (1 - ac^2 \cos^2 \varphi) \sin^2 \varphi, \\ \dot{\varphi} &= \lambda + \varepsilon (1 - ac^2 \cos^2 \varphi) \sin \varphi \cos \varphi. \end{aligned} \tag{5.32}$$

Nous avons obtenu un système de deux équations par rapport à $c$ et $\varphi$, qui est équivalent à l'équation initiale (5.28). Ce système contient une variable lente $c$ et une rapide $\varphi$. En vertu de la méthode de moyennisation, on peut déduire une solution approchée du système (5.32) en remplaçant les seconds membres du système par leurs valeurs moyennes par rapport à $\varphi$:

$$\begin{aligned} \dot{c} &= \frac{\varepsilon c}{2\pi} \int_0^{2\pi} (1 - ac^2 \cos^2 \varphi) \sin^2 \varphi \, d\varphi = \frac{\varepsilon c}{2}\left(1 - \frac{ac^2}{4}\right), \\ \dot{\varphi} &= \lambda + \frac{\varepsilon}{2\pi} \int_0^{2\pi} (1 - ac^2 \cos^2 \varphi) \sin \varphi \cos \varphi \, d\varphi = \lambda. \end{aligned} \tag{5.33}$$

A partir des équations (5.33) on peut déterminer les mouvements dont les amplitudes sont stationnaires. Il suffit à cet effet d'égaler à zéro le second membre de la première équation du système (5.33). Nous remarquons que l'équation

$$c\left(\frac{ac^2}{4} - 1\right) = 0 \tag{5.34}$$

présente trois racines: $c_1 = 0$, $c_{2,3} = \pm 2\sqrt{1/a}$, c'est-à-dire que le système de Van der Pol admet deux solutions stationnaires: l'état de repos $c = 0$ et un mouvement périodique de période $T = 2\pi/\lambda$. Nous retrouvons le résultat acquis au § 4 par la méthode de Poincaré. Mais maintenant nous sommes en mesure d'étudier ce phénomène avec bien plus de détails, d'envisager aussi le « problème d'établissement », c'est-à-dire le caractère des processus transitoires, ainsi que la *stabilité orbitale*. Ce terme qui a été introduit par Poincaré exprime la propriété des mouvements voisins de conserver une amplitude constante.

REMARQUE. Le vocable « stabilité orbitale » a été emprunté à l'astronomie. Considérons par exemple le mouvement d'un satellite de la Terre. Ce mouvement sera visiblement instable au sens de Liapounov. En effet, prenons un autre satellite qui à l'instant initial se trouve sur le rayon vecteur $r$ du premier mais décalé d'une quantité $\delta r$. Comme la période de révolution dépend du rayon, les périodes de ces satellites différeront d'une petite quantité. Ce qui fait que leur mouvement sera instable au sens de Liapounov: à chaque révolution l'écart grandira entre ces satellites. Mais étant instable, le mouvement d'un corps dans un champ de gravitation sera orbitalement stable, c'est-à-dire que les orbites de satellites voisins resteront toujours voisines.

Pour étudier la stabilité orbitale des solutions stationnaires de l'équation de Van der Pol, posons

$$c = c_i + \delta c, \tag{5.35}$$

où $c_i$ sont les racines de l'équation (5.34). Portons cette expression dans la première équation (5.33) et linéarisons l'expression obtenue par rapport à $\delta c$. Des calculs évidents nous conduisent à

$$\delta\dot{c} = \frac{\varepsilon}{2}\left[1 - \frac{3ac_i^2}{4}\right]\delta c. \tag{5.36}$$

Si $i = 1$, c'est-à-dire $c_i = 0$, alors $\delta\dot{c} = \varepsilon\delta c/2$ ou $\delta c = \delta c_0 \times$ $\times \exp\{\varepsilon t/2\}$.

Cela signifie que l'amplitude croîtra avec le temps et si à l'instant initial $\delta c > 0$, alors le point courant ne reviendra jamais à l'origine. Donc, la solution triviale de l'équation de Van der Pol est instable.

Supposons maintenant que $c_i = c_2 = 2\sqrt{1/a}$. Alors $\delta\dot{c} = -\varepsilon\delta c$, c'est-à-dire que $\delta c = \delta c_0 \exp\{-\varepsilon t\}$. Cela signifie que $\delta c \to 0$ lorsque $t$ croît. Si pour une raison ou une autre le système quitte l'orbite stationnaire, il y reviendra à la longue. S'agissant de la phase $\varphi$, elle peut varier d'une valeur aussi grande que l'on veut si seulement $\lambda$ dépend de $x$.

Certes, ces résultats ne sont pas suffisamment rigoureux. Dans nos raisonnements nous avons négligé les quantités d'ordre $O(\varepsilon^2)$, c'est-à-dire qu'on s'est placé dans le cas où $\delta c_0$ est assez petit. Ce faisant nous n'avons pas étudié complètement la stabilité au sens clas-

sique, c'est-à-dire le comportement de la solution lorsque $t \to \infty$, mais nous avons seulement estimé le comportement de la solution au voisinage des états stationnaires du système.

Signalons que cette estimation présente précisément un intérêt particulier pour la résolution de problèmes pratiques d'analyse de systèmes concrets et ce d'autant plus que pour $\delta c_0$ et $\varepsilon$ petits cette estimation est valable sur des intervalles de temps relativement grands de l'ordre de $O(1/\varepsilon)$.

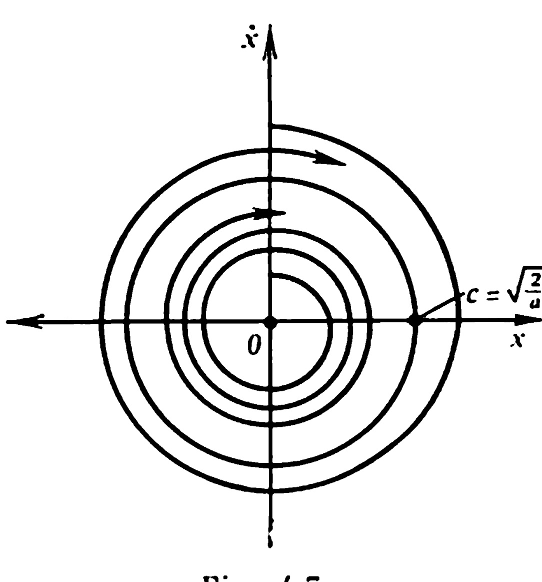


Fig. 4.7

En dépit de son caractère limité, une telle étude permet de se faire une idée de la configuration géométrique des solutions de l'équation de Van der Pol. Les trajectoires qui débutent au voisinage de l'origine des coordonnées s'éloigneront constamment de ce point : l'amplitude $c$ croîtra avec le temps. Mais elle ne pourra excéder l'amplitude stationnaire $c = \sqrt{2/a}$. Donc, les trajectoires qui commencent à l'intérieur d'une orbite stationnaire s'enrouleront sur elle de l'intérieur. Les trajectoires nées à l'extérieur d'une orbite stationnaire tendront aussi vers cet état stationnaire, c'est-à-dire s'enrouleront sur elle de l'extérieur (fig. 4.7).

La méthode qui a servi à l'analyse de l'équation de Van der Pol se généralise sans peine au cas général de systèmes à phase tournante. Considérons un système de la forme (5.1) :

$$\begin{aligned} \dot{x} &= \varepsilon X(x, y), \\ \dot{y} &= \omega(x) + \varepsilon Y(x, y). \end{aligned} \qquad (5.37)$$

Remplaçons-le par un système « tronqué », c'est-à-dire par un système d'équations moyennisées par rapport à $y$. La fréquence $\omega$ dépendant de $x$, on aura en première approximation

$$\dot{x} = \varepsilon \overline{X}(x), \qquad \dot{y} = \omega(x).$$

Les états stationnaires, c'est-à-dire les mouvements ayant une « amplitude » constante $x$, se déduisent de l'équation

$$\overline{X}(x) = 0. \qquad (5.38)$$

L'équation (5.38) est une équation transcendante en $x$. Elle peut soit ne pas posséder de racines, soit admettre un spectre discret de racines, soit être identiquement satisfaite par rapport à $x$.

Supposons qu'elle admet une racine au moins et soit $x^*$ l'une d'elles. Etudions la stabilité orbitale du mouvement d'amplitude $x^*$. Posons à cet effet $x = x^* + \delta x$ et linéarisons la première équation du système (5.37). On obtient l'équation

$$\delta \dot{x} = \varepsilon A \, \delta x, \tag{5.39}$$

où $A$ est la matrice des dérivées partielles: $A = \left(\frac{\partial \bar{X}^i}{\partial x^j}\right)$ calculées pour $x = x^*$.

L'équation (5.39) est une équation linéaire homogène à coefficients constants et l'étude du comportement de $\delta x$ se ramène au calcul des valeurs propres de la matrice $A$. En effet, on cherchera la solution de l'équation (5.39) sous la forme

$$\delta x = a e^{\varepsilon \lambda t}. \tag{5.40}$$

En portant (5.40) dans (5.39), on obtient $| A - \lambda E | = 0$. Le déterminant caractéristique

$$| A - \lambda E | \tag{5.41}$$

est un polynôme de $\lambda$; pour que la solution soit asymptotiquement stable, il est nécessaire et suffisant que les parties réelles des racines de ce polynôme soient strictement négatives. Les conditions nécessaires et suffisantes que doivent satisfaire les coefficients du polynôme (5.41) pour que ce dernier jouisse de la propriété indiquée sont exprimées par le critère de Hurwitz (cf. par exemple [67]).

On peut étudier la stabilité de la solution de l'équation (5.39) en procédant autrement. Multiplions les deux membres de l'équation (5.39) par $\delta x$. Comme

$$(\delta \dot{x}, \delta x) = \frac{1}{2} \frac{d}{dt} \delta x^2 = \frac{1}{2} \frac{d}{dt} \sum (\delta x^i)^2,$$

on peut mettre (5.39) sous la forme

$$d\rho^2/dt = 2\varepsilon \, (A \delta x, \delta x), \tag{5.42}$$

où $\rho$ désigne le rayon de la sphère des « variations »: $\rho = \sqrt{\sum (\delta x^i)^2}$.

Remarquons par ailleurs que

$$2 (A \delta x, \delta x) = \sum_{i, j} a_{ij} \delta x^i \delta x^j + \sum_{j, i} a_{ji} \delta x^i \delta x^j =$$

$$= \sum_{i, j} r_{ij} \delta x^i \delta x^j = (R \delta x, \delta x),$$

où $a_{ij}$ sont les éléments de la matrice $A$. La matrice $R$ est symétrique: $r_{ij} = a_{ij} + a_{ji}$.

Donc, l'équation (5.42) devient

$$d\rho^2/dt = \varepsilon \, (R \, \delta x, \delta x). \tag{5.43}$$

De là il s'ensuit qu'une condition suffisante pour que la solution triviale de l'équation (5.39) soit asymptotiquement stable (c'est-à-dire pour que $\rho^2 \to 0$ lorsque $t \to \infty$) est que la forme quadratique symétrique $\sum_{i,j} r_{ij}\delta x^i \delta x^j$ soit définie négative.

Souvenons-nous maintenant du critère de Sylvestre : pour qu'une forme quadratique $(R\,\delta x, \delta x)$ soit définie négative, il est nécessaire et suffisant que

$$r_{11} < 0, \quad \begin{vmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{vmatrix} > 0, \quad \begin{vmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix} < 0, \ \ldots \tag{5.44}$$

Donc, pour que les solutions de l'équation (5.39) soient asymptotiquement stables, il est suffisant que les éléments de la matrice $A$ vérifient les inégalités (5.44). Les inéquations (5.44) définissent dans l'espace des paramètres de l'équation (5.39) un domaine appelé naturellement domaine de stabilité.

L'analyse de la stabilité avec le critère de Sylvestre nous conduit exactement aux mêmes résultats qu'avec le critère de Hurwitz.

d) *Système contenant des termes oscillatoires.* Revenons maintenant au problème envisagé au § 3. On rappelle qu'on a étudié le système (3.20)

$$\dot{z} = f(z, \xi_1, \xi_2), \quad \dot{\xi}_1 = \frac{a_{12}a_{22}(z)}{a_{21}} \xi_2, \quad \dot{\xi}_2 = -\frac{a_{21}a_{11}}{a_{12}} \xi_1. \tag{5.45}$$

Le système (5.45) est intéressant pour une foule de raisons. Tout d'abord, il est le représentant d'une vaste classe de systèmes contenant un élément oscillatoire en l'absence de forces extérieures. Ensuite, la nécessité d'étudier les systèmes de la forme (5.45) apparaît lors de l'analyse numérique de « bons » systèmes, semble-t-il, dont les éléments oscillatoires ne sont pas explicités. Ces questions feront l'objet de ce numéro.

Le changement de variables (3.21)

$$\xi_1 = c\cos\varphi, \quad \dot{\xi}_1 = -\sqrt{a_{11}a_{22}}\, c\sin\varphi$$

ramène le système (5.45) au système suivant :

$$\begin{aligned} \dot{z} &= \varepsilon F_1(z, c, \varphi), \\ \dot{c} &= \varepsilon \Psi_1(z, c, \varphi)\sin^2\varphi, \\ \dot{\varphi} &= \Omega(z) + \varepsilon\Psi_2(z, c, \varphi)\sin\varphi\cos\varphi. \end{aligned} \tag{5.46}$$

Nous avons vu (au § 3) que l'application directe de la méthode du petit paramètre à l'analyse de ce système était stérile. Mais en même temps le système (5.46) n'est qu'un cas particulier du système (5.1)

dont les méthodes asymptotiques d'intégration ont été développées dans ce paragraphe. Si l'on se borne à la première approximation, c'est-à-dire à une précision de l'ordre de $O(\varepsilon)$, on peut étudier le système (5.46) de pair avec le système

$$\begin{aligned} \dot{z} &= \varepsilon \overline{F_1(z, c, \varphi)} = \varepsilon F(z, c), \\ \dot{c} &= \varepsilon \overline{\Psi_1(z, c, \varphi) \sin^2 \varphi} = \varepsilon \Psi(z, c), \\ \dot{\varphi} &= \Omega(z). \end{aligned} \tag{5.47}$$

Dans ce système les variables rapides sont séparées des lentes et l'intégration numérique est bien plus facile que pour le système primitif (5.46).

Le changement de variables et la méthode de moyennisation, utilisés pour analyser le système (5.45), peuvent être appliqués pour lever les difficultés liées à l'apparition d'oscillations de haute fréquence, que l'on rencontre souvent dans les calculs. Traitons cette question plus en détails.

Considérons un système de la forme

$$\dot{x} = \varepsilon f(x, t) \tag{5.48}$$

et résolvons le problème de Cauchy à l'aide d'une méthode numérique classique. Le paramètre $\varepsilon$ figure comme facteur multiplicatif au second membre pour souligner que la dérivée $\dot{x}$ de la variable de phase est petite en valeur absolue. Néanmoins il est possible que le pas d'intégration qui est automatiquement choisi commence à se morceler. Si l'on construit la trajectoire avec un traceur de courbes ou un display, on voit apparaître une courbe ondulée de petite longueur d'onde. Cette circonstance traduit les propriétés internes du système. Elle exprime que certaines fréquences propres du système sont élevées et bien que le processus évolue lentement — la dérivée est de l'ordre de $\varepsilon$ — il donne lieu à de petites oscillations. Même si ces oscillations ne présentent aucun intérêt pour l'analyste, celui-ci est forcé de consacrer le plus gros du temps de l'ordinateur à leur calcul. L'efficacité de la suite de l'analyse dépend de notre aptitude à mettre en évidence et à moyenniser ces oscillations à haute fréquence.

Supposons qu'à un instant $t = 0$ on connaît $x(0) = x_0$. Posons

$$x = x_0 + y \tag{5.49}$$

et mettons (5.48) sous la forme

$$\dot{y} = \varepsilon A y + \varepsilon \varphi(y), \tag{5.50}$$

où $A$ est une matrice approchée, par exemple $A = \left(\frac{\partial f}{\partial x}\right)_{x=x_0}$ et $\varphi(y) = f(x_0 + y, t) - Ay$. Les oscillations de haute fréquence ne peu-

vent être liées qu'au fait que parmi les valeurs propres de la matrice $A$ il en existe au moins deux $\lambda_1 = \alpha + i\beta$, $\lambda_2 = \alpha - i\beta$ pour lesquelles $\beta$ est de l'ordre de $O(1/\varepsilon)$.

Considérons le changement de variables linéaire

$$z = Cy \tag{5.51}$$

qui ramène la matrice $A$ à la forme normale de Jordan (on admettra par souci de simplicité que les valeurs propres $\lambda_1$ et $\lambda_2$ sont simples). Le système d'équations (5.50) s'écrit alors dans les nouvelles variables

$$\begin{aligned} \dot{z}_1 &= \varepsilon [(\alpha + i\beta) z_1 + \varphi_1^* (z_1, \ldots, z_n)], \\ \dot{z}_2 &= \varepsilon [(\alpha - i\beta) z_2 + \varphi_2^* (z_1, \ldots, z_n)], \\ \dot{z}_3 &= \varepsilon [\lambda_3 z_3 + \varphi_3^* (z_1, \ldots, z_n)], \\ &\cdots\cdots\cdots\cdots\cdots \\ \dot{z}_n &= \varepsilon [\lambda_n z_n + \varphi_n^* (z_1, \ldots, z_n)]. \end{aligned} \tag{5.52}$$

Les variables $z_i$ ne sont pas commodes à manipuler, car elles sont complexes, mais on peut facilement passer à des variables réelles. Posons en effet $\eta_1 = z_1 + z_2$, $\eta_2 = i(z_1 - z_2)$. Dans ces variables les deux premières équations du système deviennent

$$\begin{aligned} \dot{\eta}_1 &= \varepsilon [\alpha\eta_1 + \beta\eta_2 + \varphi_1^{**} (\eta_1, \eta_2, z_3, \ldots, z_n)], \\ \dot{\eta}_2 &= \varepsilon [\alpha\eta_2 - \beta\eta_1 + \varphi_2^{**} (\eta_1, \eta_2, z_3, \ldots, z_n)]. \end{aligned} \tag{5.53}$$

Donc, le système peut être ramené par des transformations linéaires à une forme qui autorise l'utilisation des variables de Van der Pol. Mais auparavant effectuons le changement $\varepsilon\beta = \omega$ et mettons le système (5.53) sous la forme

$$\dot{\eta}_1 = \omega\eta_2 + \varepsilon\Phi_1, \qquad \dot{\eta}_2 = -\omega\eta_1 + \varepsilon\Phi_2. \tag{5.54}$$

Posons ensuite

$$\eta_1 = c \cos\varphi, \qquad \dot{\eta}_1 = -c\omega \sin \varphi. \tag{5.55}$$

Poursuivons la procédure habituelle: dérivons la première équation (5.55) et égalons-la à la seconde:

$$\dot{c} \cos \varphi - c\dot{\varphi} \sin \varphi = -\omega c \sin \varphi. \tag{5.56}$$

Déterminons ensuite $\eta_2$ à partir de la première équation (5.54):

$$\eta_2 = \frac{1}{\omega} \dot{\eta}_1 - \frac{\varepsilon}{\omega} \Phi_1 = -c \sin \varphi - \frac{\varepsilon}{\omega} \Phi_1.$$

En dérivant cette expression et en l'égalant à la deuxième équation du système (5.54), ou obtient l'équation

$$-\dot{c}\sin\varphi - c\dot{\varphi}\cos\varphi = \frac{\varepsilon}{\omega}\dot{\Phi}_1 - \omega c\cos\varphi + \varepsilon\Phi_2. \qquad (5.57)$$

Résolvons le système d'équations (5.56) et (5.57) par rapport à $\dot{c}$ et $\dot{\varphi}$. On trouve

$$\begin{aligned} \dot{c} &= -\varepsilon\left(\frac{1}{\omega}\dot{\Phi}_1 + \Phi_2\right)\sin\varphi, \\ \dot{\varphi} &= \omega - \frac{\varepsilon}{c}\left(\frac{1}{\omega}\dot{\Phi}_1 + \Phi_2\right)\cos\varphi. \end{aligned} \qquad (5.58)$$

Les équations (5.58) remplacent les deux premières équations du système (5.52). Les autres équations du système peuvent être mises sous la forme

$$\dot{z}_i = \varepsilon\Phi_i, \qquad i = 3, 4, \ldots, n, \qquad (5.59)$$

où

$$\Phi_i = \Phi_i(c, \varphi, z_3, \ldots, z_n).$$

Le système d'équations (5.58), (5.59) se rapporte à la classe des systèmes (5.32) qui contiennent une variable rapide $\varphi$. La variable rapide $\varphi$ figure dans les seconds membres de ce système seulement par l'intermédiaire des fonctions trigonométriques $\cos\varphi$ et $\sin\varphi$, donc ces seconds membres sont des fonctions périodiques de $\varphi$ de période $2\pi$. On peut ainsi appliquer la méthode de moyennisation pour intégrer ce système. Ici la fréquence $\omega$ est constante, donc la première approximation sera le système d'équations

$$\begin{aligned} \dot{c} &= -\frac{\varepsilon}{2\pi}\int_0^{2\pi}\left\{\frac{1}{\omega}\dot{\Phi}_1 + \Phi_2\right\}\sin\varphi\, d\varphi = \varepsilon\Psi_1(c, z_3, \ldots, z_n), \\ \dot{\varphi} &= \omega - \frac{\varepsilon}{2\pi}\int_0^{2\pi}\left\{\frac{1}{\omega}\dot{\Phi}_1 + \Phi_2\right\}\cos\varphi\, d\varphi = \omega - \varepsilon\Psi_2(c, z_3, \ldots, z_n), \\ \dot{z}_i &= \frac{\varepsilon}{2\pi}\int_0^{2\pi}\Phi_i\, d\varphi = \varepsilon\Psi_i(c, z_3, \ldots, z_n), \\ & \qquad i = 3, 4, \ldots, n. \end{aligned} \qquad (5.60)$$

La variable rapide $\varphi$ est exclue des seconds membres du système (5.58), (5.59) et le pas d'intégration peut être pris grand.

Certes, le passage du système primitif (5.48) au système (5.60) suppose la réalisation d'une série de transformations qui, en principe,

sont complexes et réclament un certain temps machine. Mais le système obtenu peut être tellement simple que cette perte de temps machine se compense totalement. Les techniciens du Centre de calcul de l'Académie des sciences d'U.R.S.S. ont traité des cas dans lesquels le passage aux systèmes moyennisés leur a permis d'accroître le pas d'intégration de plusieurs milliers de fois. Il s'en est suivi non seulement un gain de temps mais aussi la disparition de l'imprévisible erreur de calcul liée aux instruments et qui est inévitable dans les longs calculs.

Pour mettre en évidence les éléments oscillatoires du système (5.50), nous avons ramené le système linéaire à la forme de Jordan. Dans les problèmes pratiques, il faut essayer d'éviter cette procédure car elle conduit à des valeurs complexes et implique une opération supplémentaire pour passer à des variables réelles. De plus la réduction à la forme de Jordan réclame le calcul de toutes les nouvelles variables, or seule la composante oscillatoire principale nous intéresse. On peut donc tenter de la dégager en faisant jouer les particularités physiques du système.

On pourrait proposer d'autres méthodes. Supposons par exemple que la valeur propre de plus grand module est imaginaire pure. On peut alors introduire une nouvelle variable scalaire $z$ qui est liée à la variable $x$ de l'équation

$$\dot{x} = Ax \tag{5.61}$$

par les transformations linéaires

$$z = (b,\ x), \qquad \dot{z} = (c,\ x). \tag{5.62}$$

Exigeons que la nouvelle variable $z$ soit solution de l'équation différentielle ordinaire du second ordre

$$\ddot{z} + \omega^2 z = 0. \tag{5.63}$$

En dérivant la première relation (5.62) et en comparant avec la seconde, on obtient

$$(b,\ Ax) = (c,\ x),$$

d'où

$$(x,\ A^*b - c) = 0. \tag{5.64}$$

L'équation (5.64) devant être vérifiée pour tout $x$, on a

$$c = A^*b. \tag{5.65}$$

En dérivant ensuite la deuxième relation (5.62) et en portant dans (5.63), on trouve $(c,\ Ax) + \omega^2\ (b,\ x) = 0$, ou $(x,\ A^*c - \omega^2 b) =$ $= 0$, d'où l'on déduit immédiatement une autre relation entre les

vecteurs $c$ et $b$:

$$A^*c - \omega^2 b = 0.$$

L'égalité (5.65) aidant, on obtient en définitive

$$[(A^*)^2 - \omega^2 E]\, b = 0, \tag{5.66}$$

où $E$ est la matrice unité. Pour que l'équation (5.66) admette une solution non triviale, il est nécessaire et suffisant que $\omega^2$ soit racine de l'équation

$$|\, (A^*)^2 - \omega^2 E \,| = 0, \tag{5.67}$$

plus exactement, la racine de plus grand module.

Une fois $\omega$ connue, on trouve $b$ et ensuite le vecteur $c$ à l'aide de la formule (5.65) et enfin la relation entre $x$ et $z$ à l'aide des formules (5.62). Du système (5.61) on élimine ensuite deux équations quelconques, par exemple des équations contenant $\dot{x}_1$ et $\dot{x}_2$, on lui adjoint l'équation (5.63) et on détermine les variables $x_1$ et $x_2$ à partir des relations (5.62).

Ce schéma peut être modifié ou simplifié de diverses manières en fonction des particularités du système. Signalons enfin qu'il n'est point besoin de connaître la valeur exacte de $\omega$, il faut simplement d'établir une équation de la forme (5.63) et une relation de la forme (5.62).

Une remarque pratique. La simplification du système (5.48) relevait de l'analyse de la matrice $A$ engendrée par la transformation (5.49). Donc, nous avons mis en évidence l'élément oscillatoire par une analyse des seconds membres du système à un instant fixe. Mais les propriétés du système varient avec le temps. En commençant l'intégration du système (5.60) avec un pas élevé, on risque ensuite d'être confronté à un fractionnement du pas et à l'apparition de nouvelles oscillations. Dans ces conditions il faut stopper les calculs et reprendre la procédure avec une nouvelle matrice $A$ déterminée avec d'autres valeurs de la variable de phase.

e) *Systèmes contenant des éléments giratoires.* Voyons encore une classe de systèmes qui se prêtent bien à l'analyse par la méthode de moyennisation. Soit le système

$$\dot{\xi} = R(\xi, z), \qquad \ddot{z} + f(z, \xi) = 0, \tag{5.68}$$

où $\xi$ est un vecteur de dimension $N$, $z$, un scalaire. On admettra que pour tout $\xi$ la fonction $f(z, \xi)$ est une fonction périodique en $z$ de période donnée $T$, et de plus

$$f(\xi) = \frac{1}{T}\int_0^T f(z, \xi)\, dz = 0. \tag{5.69}$$

On supposera aussi que

$$\dot{z}(0) = \Omega \tag{5.70}$$

est un grand nombre.

Dans ces conditions le système (5.68) sera appelé *système à élément giratoire*. Ces systèmes se rencontrent fréquemment dans les applications. Ainsi, le problème général du mouvement d'un satellite, c'est-à-dire le problème de l'étude simultanée du mouvement de

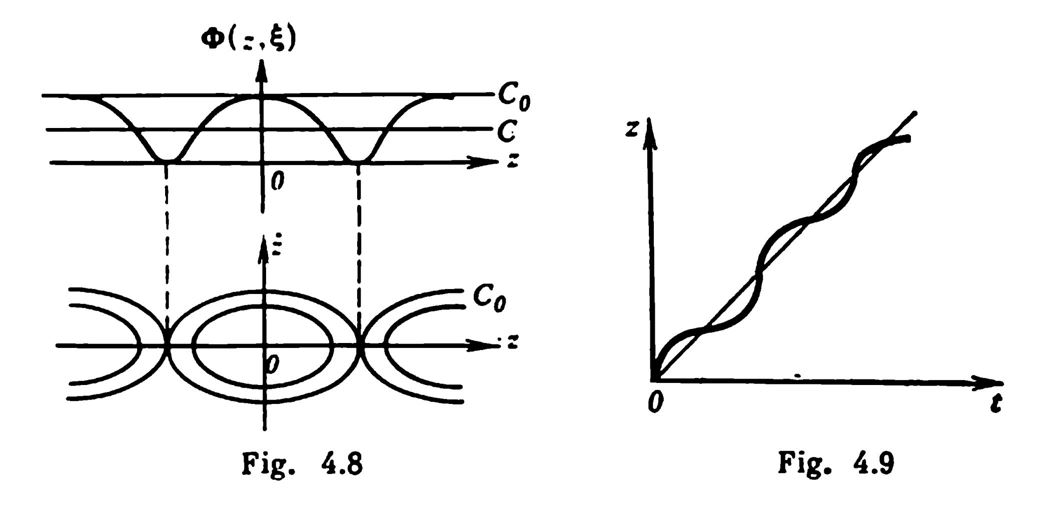



Fig. 4.8 Fig. 4.9

son centre d'inertie et du mouvement par rapport à ce centre d'inertie se ramène dans nombre de cas à un système à élément giratoire.

Fixons $\xi$ et considérons le mouvement décrit par la deuxième équation du système (5.68). Multiplions cette équation par $\dot{z}$:

$$\ddot{z}\dot{z} + \dot{z}f(z, \xi) = 0,$$

ou

$$\frac{1}{2}\frac{d}{dt}\dot{z}^2 + \frac{d}{dt}\int_0^z f(\alpha, \xi)\, d\alpha = 0.$$

Posons

$$\Phi(z, \xi) = \int_0^z f(\alpha, \xi)\, d\alpha.$$

En se servant de la propriété (5.69), on obtient $\Phi(0, \xi) = \Phi(T, \xi) = \Phi(-T, \xi)$. Donc

$$\dot{z} = \pm\sqrt{-2\Phi(z, \xi) + C}, \tag{5.71}$$

où $C$ est une constante arbitraire caractérisant l'énergie du système. La configuration des trajectoires de ce système est représentée sur la figure 4.8. Aux petites valeurs de l'énergie $C$ correspondent des mou-

vements oscillatoires fermés. Si $C > C_0$, alors pour tout $\xi$, le système accomplit des rotations. Sa coordonnée $z$ croît indéfiniment avec le temps. Pour les grandes valeurs de $C$ et à $\xi$ fixe, la fonction $z = z(t, \xi)$ est représentée graphiquement sur la figure 4.9.

Pour les petites valeurs de l'énergie, $z$ décrit un mouvement oscillatoire et elle peut être éliminée du système par les méthodes développées dans ce paragraphe. On voit sur la figure 4.9 que l'élément $z$ oscille et croît en même temps. Donc, pour utiliser la technique de moyennisation, il faut une approche différente de celle qui a servi à analyser les systèmes à élément oscillant.

On admettra donc que la quantité $\Omega$ définie par (5.70) est grande. Alors en désignant $\varepsilon = 1/\Omega$, en faisant le changement de variable $t = \varepsilon s$ et en posant de plus $\dot{z} = \Omega + x$, on réduit le système (5.68) à la forme

$$\frac{d\xi}{ds} = \varepsilon R(\xi, z), \quad \frac{dx}{ds} = -\varepsilon f(\xi, z), \quad \frac{dz}{ds} = 1 + \varepsilon x. \tag{5.72}$$

Si $R$ est une fonction périodique de la variable rapide $z$, alors le système (5.72) est un cas particulier des systèmes à phase tournante et l'on peut se servir de la technique de moyennisation classique. A noter qu'on obtiendrait en première approximation

$$\frac{d\xi}{ds} = \varepsilon \overline{R}(\xi), \quad \frac{dx}{ds} = 0, \quad \frac{dz}{ds} = 1, \tag{5.73}$$

où

$$\overline{R}(\xi) = \frac{1}{T} \int_0^T R(\xi, z)\, dz,$$

c'est-à-dire que le mouvement étudié est voisin d'une rotation uniforme. Si la fonction $R$ est une fonction de $z$ quasi périodique ou oscillatoire bornée, alors la technique de moyennisation passe aussi, mais la moyennisation sur la période $T$ doit être remplacée par le calcul d'intégrales de la forme

$$\overline{R}(\xi) = \lim_{T \to \infty} \frac{1}{T} \int_0^T R(\xi, z)\, dz.$$

## § 6. Cas de plusieurs degrés de liberté oscillatoires

Poursuivons l'étude des systèmes de la forme

$$\dot{x} = \varepsilon X(x, y, \varepsilon), \qquad \dot{y} = \omega(x) + \varepsilon Y(x, y, \varepsilon), \tag{6.1}$$

où $X$ et $Y$ sont des fonctions périodiques de $y$. Dans le paragraphe précédent, nous avons traité le cas où la variable $y$ était une variable

scalaire. Nous avons vu que ce cas décrivait un système à un seul degré de liberté oscillant rapidement. Nous avons réussi par la transformation de Van der Pol à réduire ce système à la forme (6.1), où $y$ est un scalaire.

Si le système contient plusieurs éléments oscillatoires ou giratoires, on peut encore utiliser la transformation de Van der Pol (dans le cas d'éléments oscillatoires) ou une transformation identique à celle qui a servi à étudier les systèmes à élément giratoire, et réduire le système primitif à un système de la forme (6.1).

Le passage à plusieurs phases tournantes, c'est-à-dire à des systèmes dont plusieurs degrés de liberté sont animés d'oscillations, n'est pas trivial et il n'est pas toujours possible d'étendre directement les résultats du paragraphe précédent au cas étudié. Le cas où la variable rapide est un vecteur réclame un appareil spécial. Mais avant de passer à son examen, signalons que de très nombreux problèmes se ramènent à l'étude de systèmes de la forme (6.1) dans lesquels le vecteur $y$ est de dimension $>1$. Le plus élémentaire d'entre eux est probablement le problème du pendule à caractéristique quasi linéaire soumis à l'action d'une force périodique:

$$\ddot{z} + \omega^2 z = \varepsilon F(z, \dot{z}, t), \tag{6.2}$$

où $F$ est une fonction périodique de $t$ de période $2\pi$. En utilisant la transformation de Van der Pol

$$z = x \cos y, \qquad \dot{z} = -x\omega \sin y$$

et en admettant que $dt/d\tau = 1$, on ramène l'équation (6.2) à la forme

$$\frac{dx}{d\tau} = \varepsilon X(x, y, t), \quad \frac{dy}{d\tau} = \omega + \varepsilon Y(x, y, t), \quad \frac{dt}{d\tau} = 1. \tag{6.3}$$

Le système (6.3) est un cas particulier du système (6.1), puisque ses seconds membres sont des fonctions périodiques de $y$ et $t$ de même période $2\pi$. Ces situations se présentent aussi bien en mécanique, qu'en écologie, en économie, etc.

Passons maintenant à l'étude des caractéristiques du système (6.1).

a) *Systèmes à deux variables rapides.* On se bornera au cas où le vecteur $y$ est de dimension deux. Les principales caractéristiques liées au passage du cas scalaire au cas vectoriel peuvent être étudiées sur l'exemple d'un système à deux phases tournantes.

Soit donc le système

$$\begin{aligned} \dot{x} &= \varepsilon X(x, y, z, \varepsilon), \\ \dot{y} &= \omega(x) + \varepsilon Y(x, y, z, \varepsilon), \\ \dot{z} &= \lambda(x) + \varepsilon Z(x, y, z, \varepsilon), \end{aligned} \tag{6.4}$$

où $X$, $Y$ et $Z$ sont des fonctions périodiques des variables scalaires $y$ et $z$, de périodes respectives $T_y = 2\pi/l$ et $T_z = 2\pi/m$.

Essayons en suivant la même démarche qu'avant de trouver une transformation qui permette de séparer les variables lentes des rapides et de ramener le calcul des variables rapides à des intégrations. Posons

$$\begin{aligned} x &= \bar{x} + \varepsilon u(\bar{x}, \bar{y}, \bar{z}, \varepsilon), \\ y &= \bar{y} + \varepsilon v(\bar{x}, \bar{y}, \bar{z}, \varepsilon), \\ z &= \bar{z} + \varepsilon w(\bar{x}, \bar{y}, \bar{z}, \varepsilon), \end{aligned} \tag{6.5}$$

où les nouvelles variables sont solutions du système d'équations

$$\begin{aligned} \dot{\bar{x}} &= \varepsilon A(\bar{x}, \varepsilon), \\ \dot{\bar{y}} &= \omega(\bar{x}) + \varepsilon B(\bar{x}, \varepsilon), \\ \dot{\bar{z}} &= \lambda(\bar{x}) + \varepsilon C(\bar{x}, \varepsilon), \end{aligned} \tag{6.6}$$

où les fonctions des seconds membres sont inconnues *a priori*. En portant les expressions (6.5) et (6.6) dans le système d'équations (6.4) on obtient le système d'équations suivant:

$$\begin{aligned} \frac{\partial u}{\partial \bar{y}} \omega(\bar{x}) + \frac{\partial u}{\partial \bar{z}} \lambda(\bar{x}) &= X(\bar{x}, \bar{y}, \bar{z}, \varepsilon) - A(\bar{x}, \varepsilon) - \\ &- \varepsilon \frac{\partial u}{\partial \bar{x}} A(\bar{x}, \varepsilon) - \varepsilon \frac{\partial u}{\partial \bar{y}} B(\bar{x}, \varepsilon) - \varepsilon \frac{\partial u}{\partial \bar{z}} C(\bar{x}, \varepsilon) + \\ &+ \{X(\bar{x} + \varepsilon u, \bar{y} + \varepsilon v, \bar{z} + \varepsilon w, \varepsilon) - X(\bar{x}, \bar{y}, \bar{z}, \varepsilon)\}, \\ \frac{\partial v}{\partial \bar{y}} \omega(\bar{x}) + \frac{\partial v}{\partial \bar{z}} \lambda(\bar{x}) &= \frac{\omega(\bar{x} + \varepsilon u) - \omega(\bar{x})}{\varepsilon} + Y(\bar{x}, \bar{y}, \bar{z}, \varepsilon) - \\ &- B(\bar{x}, \varepsilon) - \varepsilon \frac{\partial v}{\partial \bar{x}} A(\bar{x}, \varepsilon) - \varepsilon \frac{\partial v}{\partial \bar{y}} B(\bar{x}, \varepsilon) - \varepsilon \frac{\partial v}{\partial \bar{z}} C(\bar{x}, \varepsilon) + \\ &+ \{Y(\bar{x} + \varepsilon u, \bar{y} + \varepsilon v, \bar{z} + \varepsilon w, \varepsilon) - Y(\bar{x}, \bar{y}, \bar{z}, \varepsilon)\}, \qquad (6.7) \\ \frac{\partial w}{\partial \bar{y}} \omega(\bar{x}) + \frac{\partial w}{\partial \bar{z}} \lambda(\bar{x}) &= \frac{\lambda(\bar{x} + \varepsilon u) - \lambda(\bar{x})}{\varepsilon} + Z(\bar{x}, \bar{y}, \bar{z}, \varepsilon) - \\ &- C(\bar{x}, \varepsilon) - \varepsilon \frac{\partial w}{\partial \bar{x}} A(\bar{x}, \varepsilon) - \varepsilon \frac{\partial w}{\partial \bar{y}} B(\bar{x}, \varepsilon) - \varepsilon \frac{\partial w}{\partial \bar{z}} C(\bar{x}, \varepsilon) + \\ &+ \{Z(\bar{x} + \varepsilon u, \bar{y} + \varepsilon v, \bar{z} + \varepsilon w, \varepsilon) - Z(\bar{x}, \bar{y}, \bar{z}, \varepsilon)\}. \end{aligned}$$

Comme dans le paragraphe précédent, étudions le système (6.7) par la méthode des approximations successives. Prenons pour pre-

mière approximation le système suivant

$$\begin{aligned} \frac{\partial u}{\partial \bar{y}} \omega(\bar{x}) + \frac{\partial u}{\partial \bar{z}} \lambda(\bar{x}) &= X(\bar{x}, \bar{y}, \bar{z}, \varepsilon) - A(\bar{x}, \varepsilon), \\ \frac{\partial v}{\partial \bar{y}} \omega(\bar{x}) + \frac{\partial v}{\partial \bar{z}} \lambda(\bar{x}) &= Y(\bar{x}, \bar{y}, \bar{z}, \varepsilon) - B(\bar{x}, \varepsilon), \\ \frac{\partial w}{\partial \bar{y}} \omega(\bar{x}) + \frac{\partial w}{\partial \bar{z}} \lambda(\bar{x}) &= Z(\bar{x}, \bar{y}, \bar{z}, \varepsilon) - C(\bar{x}, \varepsilon). \end{aligned} \tag{6.8}$$

Les approximations suivantes auront une structure identique. Dans ce paragraphe on se limitera à la première approximation. Comme dans le cas scalaire traité au paragraphe précédent, la séparation des mouvements lents et rapides se ramène à l'intégration d un système d'équations aux dérivées partielles du premier ordre. Mais dans le cas d'une variable rapide nous avons obtenu un système d'équations résolues relativement à la dérivée par rapport à la variable rapide. Les équations ne renfermaient pas d'autres dérivées. Ceci nous a permis d'acquérir la solution par une intégration directe. Nous avons maintenant affaire à deux dérivées, l'une par rapport à $\bar{y}$, l'autre par rapport à $\bar{z}$ et la méthode du paragraphe précédent est inadéquate.

Les seconds membres des équations (6.8) étant des fonctions périodiques des variables rapides, la méthode de Fourier semble tout indiquée à leur analyse.

Outre les fonctions inconnues $u(\bar{x}, \bar{y}, \bar{z}, \varepsilon)$, $v(\bar{x}, \bar{y}, \bar{z}, \varepsilon)$ et $w(\bar{x}, \bar{y}, \bar{z}, \varepsilon)$ les équations (6.8) contiennent des fonctions $A$, $B$ et $C$ qui restent à déterminer. Pour le faire, on exigera, comme au paragraphe précédent, que les fonctions $u$, $v$ et $w$ soient bornées lorsque $\bar{y} \to \infty$ et $\bar{z} \to \infty$.

b) *Méthode de Fourier*. Elle repose sur la possibilité de représenter la fonction périodique $f(y)$ de période $T = 2\pi/l$ par la série

$$f(y) = \sum_{k=-\infty}^{k=+\infty} f_k e^{ikly},$$

où les coefficients $f_k$ sont définis par la formule

$$f_k = \frac{1}{T} \int_0^T f(y) e^{-ikly} dy, \tag{6.9}$$

où $i = \sqrt{-1}$. Si la fonction $f$ est une fonction de deux variables: $f = f(y, z)$, périodique, de période $T_y = 2\pi/l$ en $y$ et $T_z = 2\pi/m$, en $z$, alors elle se développe en une double série

$$f(y, z) = \sum_{k=-\infty}^{k=+\infty} \sum_{s=-\infty}^{s=+\infty} f_{ks} e^{i(kly+smz)}.$$

Considérons la première équation du système (6.8). La fonction $X(\bar{x}, \bar{y}, \bar{z}, \varepsilon)$ est une fonction périodique en $\bar{y}$ et $\bar{z}$ de périodes respectives $T_y$ et $T_z$. Nous pouvons donc la mettre sous la forme

$$X(\bar{x}, \bar{y}, \bar{z}, \varepsilon) = \sum_{k,\, s=-\infty}^{k,\, s=+\infty} a_{ks}^{(x)}(\bar{x}, \varepsilon)\, e^{i(kl\bar{y}+sm\bar{z})},$$

où les coefficients du développement $a_{ks}^{(x)}(\bar{x}, \varepsilon)$ sont définis par des formules identiques aux formules (6.9):

$$a_{ks}^{(x)}(\bar{x}, \varepsilon) = \frac{1}{T_y T_z} \int_0^{T_y} \int_0^{T_z} X(\bar{x}, \bar{y}, \bar{z}, \varepsilon)\, e^{-i(kl\bar{y}+sm\bar{z})}\, d\bar{y}\, d\bar{z}. \qquad (6.9')$$

Donc, la première équation (6.8) peut être récrite sous la forme

$$\frac{\partial u}{\partial \bar{y}} \omega + \frac{\partial u}{\partial \bar{z}} \lambda = \sum_{k,\, s} a_{ks}^{(x)}(\bar{x}, \varepsilon)\, e^{i(kl\bar{y}+sm\bar{z})} + a_{00} - A(\bar{x}, \varepsilon). \qquad (6.10)$$

Dans (6.10) la sommation est étendue à tous les indices $k$, $s$ non simultanément nuls. Le coefficient $a_{00}$ se calcule par la formule

$$a_{00}^{(x)}(\bar{x}, \varepsilon) = \frac{1}{T_y T_z} \int_0^{T_y} \int_0^{T_z} X(\bar{x}, \bar{y}, \bar{z}, \varepsilon)\, d\bar{y}\, d\bar{z}. \qquad (6.9'')$$

Etant donné que le terme $a_{00} - A(\bar{x}, \varepsilon)$ du second membre de l'équation (6.10) ne contient pas d'harmonique, la fonction dont la dérivée est égale au second membre de l'équation (6.10) doit être la somme d'une fonction périodique en ses deux variables $\bar{y}$ et $\bar{z}$ et d'une fonction linéaire $c\bar{y} + d\bar{z}$, dont la dérivée est égale à $a_{00} - A(\bar{x}, \varepsilon)$. Autrement dit, la fonction $u(\bar{x}, \bar{y}, \bar{z}, \varepsilon)$ doit se représenter par la série

$$u(\bar{x}, \bar{y}, \bar{z}, \varepsilon) = \sum_{k=-\infty}^{k=+\infty} \sum_{s=-\infty}^{s=+\infty} b_{ks}(\bar{x}, \varepsilon)\, e^{i(kl\bar{y}+sm\bar{z})} + \\ + c(\bar{x}, \varepsilon)\, \bar{y} + d(\bar{x}, \varepsilon)\, \bar{z}. \qquad (6.11)$$

Dans le second membre de cette équation, la sommation est étendue à tous les indices non simultanément nuls. En portant l'expression (6.11) dans l'équation (6.10) et en identifiant les coefficients des exponentielles de même puissance, on obtient les formules

$$b_{ks}(\bar{x}, \varepsilon) = \frac{a_{ks}^{(x)}(\bar{x}, \varepsilon)}{i(kl\omega + sm\lambda)}, \qquad (6.12)$$

$$c(\bar{x}, \varepsilon)\, \omega(\bar{x}) + d(\bar{x}, \varepsilon)\, \lambda(\bar{x}) = a_{00}(\bar{x}, \varepsilon) - A(\bar{x}, \varepsilon). \qquad (6.13)$$

La quantité $b_{00}(\bar{x}, \varepsilon)$ reste inconnue. Pour que $u$ soit borné, il est nécessaire et suffisant, ainsi qu'il résulte de l'expression (6.11), que $c(\bar{x}, \varepsilon)$ et $d(\bar{x}, \varepsilon)$ soient nuls. Mais ceci n'aura lieu que si et seulement si $A(\bar{x}, \varepsilon) = a_{00}(\bar{x}, \varepsilon)$. La formule (6.9″) aidant, on trouve immédiatement que $A(\bar{x}, \varepsilon) = \bar{X}(\bar{x}, \varepsilon)$, où la moyennisation a été effectuée sur les deux variables rapides:

$$\bar{X}(\bar{x}, \varepsilon) = \frac{1}{T_y T_z} \int_0^{T_y} \int_0^{T_z} X(\bar{x}, \bar{y}, \bar{z}, \varepsilon)\, d\bar{y}\, d\bar{z}. \tag{6.14}$$

Si nous voulons que les variables $x$ et $\bar{x}$ soient solutions d'un même problème de Cauchy, il nous faut prendre la quantité inconnue $b_{00}(\bar{x}, \varepsilon)$ égale à 0.

Les autres équations du système (6.8) se traitent de façon analogue.

Ainsi, l'analyse de la première approximation nous donne

$$\begin{gathered} A(\bar{x}, \varepsilon) = \bar{X}, \\ B(\bar{x}, \varepsilon) = \bar{Y} + \omega_x u(\bar{x}, \bar{y}, \bar{z}, \varepsilon), \\ C(\bar{x}, \varepsilon) = \bar{Z} + \lambda_x v(\bar{x}, \bar{y}, \bar{z}, \varepsilon), \\ u = \sum_{k,\, s \neq 0} \frac{a_{ks}^{(x)}(\bar{x}, \varepsilon)}{i(kl\omega + sm\lambda)} \exp i(kl\bar{y} + sm\bar{z}) + a_0^{(x)}(\bar{x}), \\ v = \sum_{k,\, s \neq 0} \frac{a_{ks}^{(y)}(\bar{x}, \varepsilon)}{i(kl\omega + sm\lambda)} \exp i(kl\bar{y} + sm\bar{z}) + a_0^{(y)}(\bar{x}), \\ w = \sum_{k,\, s \neq 0} \frac{a_{ks}^{(z)}(\bar{x}, \varepsilon)}{i(kl\omega + sm\lambda)} \exp i(kl\bar{y} + sm\bar{z}) + a_0^{(z)}(\bar{x}). \end{gathered} \tag{6.15}$$

Les formules (6.15) sont identiques à celles acquises dans le cas d'une seule variable rapide. Mais ces formules n'ont un sens que si et seulement si l'un des dénominateurs n'est pas nul:

$$kl\omega + sm\lambda \neq 0. \tag{6.16}$$

Signalons que dans le cas général $\omega$ et $\lambda$ sont fonctions de la variable lente $\bar{x}$. Donc, si la condition (6.16) ne sera pas violée pendant toute la durée du processus, l'existence de la deuxième variable rapide n'apporte aucune complication: nous devons conduire la moyennisation sur les deux variables rapides. Nous conviendrons d'appeler ce cas, *cas non résonnant*.

c) ***Etude de la résonance principale.*** On appellera résonnante la situation qui correspond au cas où l'« amplitude » $x$ se trouve au

voisinage d'une racine de l'équation

$$kl\omega(x) + sm\lambda(x) = 0. \qquad (6.17)$$

Si les « fréquences » $\omega$ et $\lambda$ sont indépendantes de $x$, alors on appellera *résonance* le phénomène qui apparaît dans le système lorsque $\omega$ et $\lambda$ sont reliées par la condition $kl\omega + sm\lambda = 0$, où $k$ et $s$ sont des entiers quelconques, strictement positifs ou négatifs. Si par ailleurs $|k| = |s|$, c'est-à-dire

$$l\omega \pm m\lambda = 0, \qquad (6.18)$$

alors ce cas sera qualifié de *résonance principale*.

L'étude des phénomènes résonnants se heurte à des difficultés mathématiques de taille. Dans cet ouvrage, on se bornera au cas élémentaire de résonance principale où $\omega$ et $\lambda$ ne dépendent pas de la variable lente $x$.

La quantité $h^* = kl\omega + sm\lambda$ s'appelle généralement *désaccord*. On conviendra de désigner par *voisinage* de la résonance une relation entre les paramètres telle que le désaccord soit petit: $h^* = \varepsilon h$. Au voisinage de la résonance on peut également procéder à une intégration asymptotique. Mais dans ce cas le comportement asymptotique sera différent de celui décrit par les formules (6.15).

Posons donc $\lambda = l\omega/m + \varepsilon h/m$ et mettons le système (6.4) sous la forme

$$\begin{aligned} \dot{x} &= \varepsilon X(x, y, z, \varepsilon), \\ \dot{y} &= \omega + \varepsilon Y(x, y, z, \varepsilon), \\ \dot{z} &= \frac{l\omega}{m} + \varepsilon\left[Z(x, y, z, \varepsilon) + \frac{h}{m}\right]. \end{aligned} \qquad (6.19)$$

On rappelle que $X$, $Y$ et $Z$ sont des fonctions périodiques de $y$ et $z$ de périodes respectives $2\pi/l$ et $2\pi/m$.

Remplaçons la variable $y$ par la variable $\theta$ appelée *décalage de phase*:

$$\theta = \frac{m}{l} z - y.$$

Utilisons le système (6.19) pour composer l'équation

$$\dot{\theta} = \varepsilon\vartheta\left(x, \frac{m}{l} z - \theta, z, \varepsilon\right), \qquad (6.20)$$

où $\vartheta$ est de la forme

$$\vartheta\left(x, \frac{m}{l} z - \theta, z, \varepsilon\right) = \left\{\left[Z\left(x, \frac{m}{l} z - \theta, z, \varepsilon\right) + \frac{h}{m}\right]\frac{m}{l} - Y\left(x, \frac{m}{l} z - \theta, z, \varepsilon\right)\right\}.$$

Mettons les autres équations du système (6.19) sous la forme suivante:

$$\dot{x} = \varepsilon X\left(x, \frac{m}{l}z - \theta, z, \varepsilon\right),$$
$$\dot{z} = \frac{l\omega}{m} + \varepsilon\left[Z\left(x, \frac{m}{l}z - \theta, z, \varepsilon\right) + \frac{h}{m}\right].$$

Le système (6.19) devient maintenant

$$\begin{aligned} \dot{x} &= \varepsilon X^*(x, \theta, z, \varepsilon),\\ \dot{\theta} &= \varepsilon\vartheta^*(x, \theta, z, \varepsilon),\\ \dot{z} &= \frac{l\omega}{m} + \varepsilon\left[Z^*(x, \theta, z, \varepsilon) - \frac{h}{m}\right], \end{aligned} \tag{6.21}$$

où $X^*$, $\vartheta^*$ et $Z^*$ ont une signification évidente. Les fonctions $X^*$ $\vartheta^*$ et $Z^*$ sont des fonctions périodiques de $z$. La variable $z$ figure de deux manières dans les seconds membres du système (6.21): seule et par l'intermédiaire de la combinaison $(m/l)\, z - \theta$. Les seconds membres du système (6.21) sont des fonctions périodiques et de $z$ avec une période $T_z = 2\pi/m$, et de la combinaison $(m/l)\, z - \theta$ avec une période $T_y = 2\pi/l$. Donc, ils sont périodiques de la combinaison $(m/l)\, z - \theta$ et de période

$$T'_z = \frac{l}{m} T_y = \frac{2\pi}{m}.$$

Ainsi, les fonctions $X^*$, $\vartheta^*$ et $Z^*$ traitées comme des fonctions de $z$, sont de période $2\pi/m$.

La quantité scalaire $\theta$ est une variable lente. Donc, contrairement au système primitif (6.4) qui est à deux phases tournantes, le système (6.21) est à une seule et par conséquent son analyse peut être menée par les méthodes de la théorie générale développée dans le paragraphe précédent. Si l'on se borne à la première approximation, on aura

$$\begin{aligned} x &= \bar{x}, & \dot{\bar{x}} &= \varepsilon\bar{X}^*(\bar{x}, \bar{\theta}, \varepsilon),\\ \theta &= \bar{\theta}, & \dot{\bar{\theta}} &= \varepsilon\bar{\vartheta}^*(\bar{x}, \bar{\theta}, \varepsilon). \end{aligned} \tag{6.22}$$

Pour déterminer la variable rapide, il faut effectuer l'intégration suivante:

$$z = z_0 + \lambda t + \varepsilon \int_0^t \bar{Z}^*(\bar{x}, \bar{\theta}, z, \varepsilon)\, dt.$$

Dans ces formules, la moyennisation est réalisée sur la période $T_2$, c'est-à-dire que

$$\overline{X}^* (x, \theta, \varepsilon) = \frac{m}{2\pi} \int_0^{2\pi/m} X^* (x, \theta, z, \varepsilon)\, dz$$

et ainsi de suite.

Ceci achève l'analyse. Nous avons exhibé un système de procédures permettant dans le cas d'une résonance principale, c'est-à-dire dans le cas où les fréquences $\omega$ et $\lambda$ sont reliées par la condition

$$l\omega \pm m\lambda = O(\varepsilon), \tag{6.23}$$

d'abaisser l'ordre du système d'une unité et de remplacer le système primitif d'équations par un système dont les solutions sont à variation lente. La méthode développée se généralise automatiquement au cas où les quantités $\omega$ et $\lambda$ sont reliées par la relation

$$kl\omega \pm sm\lambda = O(\varepsilon), \tag{6.24}$$

où $k$ et $s$ sont des entiers arbitraires.

Si $\omega$ et $\lambda$ sont des fonctions de $x$, alors les procédures d'intégration se compliquent un peu, puisque les relations (6.23) ou (6.24) peuvent être violées lorsque $x$ varie. Cela veut dire que la théorie formelle développée ici n'est valable qu'au voisinage de certaines valeurs spéciales de $x$.

* * *

La théorie développée dans ce chapitre est l'un des plus efficaces instruments de simplification des systèmes. Nous avons traité le cas où le système étudié était un système d'équations différentielles. Les systèmes décrits par des équations aux différences qui ne sont pas des approximations aux différences d'équations différentielles sont importants aussi. La généralisation des méthodes citées aux équations aux différences présente un grand intérêt pratique et théorique.

CHAPITRE V

# THÉORIE DES SYSTÈMES TIKHONOVIENS

## § 1. Considérations générales

On a déjà signalé au chapitre précédent que la simplification d'un système et la représentation du processus étudié par un modèle facilement réalisable sur ordinateur était un problème clef de l'analyse des systèmes.

Les difficultés soulevées par l'analyse numérique des modèles sur ordinateur ne sont pas seulement affaire de dimension. L'organisation des calculs risque d'être compliquée par l'apparition de processus à petits temps caractéristiques, c'est-à-dire des processus décrits par des variables rapides. Très souvent l'analyse de tels modèles implique des calculs astronomiques. C'est dire l'importance que revêt la simplification des modèles.

Mais la simplification du modèle, la substitution d'un système d'équations à un autre ne doivent pas altérer les propriétés du modèle : celui-ci doit rester fidèle au système réel. Certes toute description inexacte donne lieu à des erreurs inévitables. Mais ces erreurs doivent être « acceptables » pour les objectifs poursuivis et la précision désirée.

Le problème de la précision est loin d'être simple. En effet, toute simplification du modèle s'accompagne d'erreurs de calcul supplémentaires, y compris le « bruit » incontrôlable du processus de calcul. Les erreurs introduites en simplifiant la description du système peuvent être compensées par la réduction des erreurs consécutive à une diminution du volume des calculs. L'analyse numérique des grands systèmes nous fournit de nombreux exemples de cette nature.

Une voie d'analyse préliminaire des modèles consiste à étudier leurs propriétés asymptotiques par rapport à telle ou telle variable du modèle. Au chapitre précédent, nous avons étudié des systèmes

$$x = f(x, t, \varepsilon)$$

dont les paramètres étaient réguliers, c'est-à-dire que la fonction $f$ était une fonction analytique du paramètre $\varepsilon$ pour des valeurs assez petites de $\varepsilon$, ou, à la rigueur, continue au voisinage du point $\varepsilon = 0$, autrement dit on a toujours admis que

$$\lim_{\varepsilon \to 0} f(x, t, \varepsilon) = f(x, t, 0).$$

Nous avons appliqué les méthodes d'analyse de tels systèmes pour $\varepsilon \to 0$ à de nombreux systèmes, et notamment à des systèmes à éléments oscillatoires. Les méthodes développées au chapitre précédent permettent d'exclure ces termes oscillatoires sans altérer les propriétés fondamentales des processus étudiés et de réduire sensiblement le volume des calculs.

La classe de systèmes considérée possédait une importante propriété. Pour $\varepsilon = 0$, le système se simplifiait, parfois il se décomposait en plusieurs équations, mais il conservait son ordre et le problème de Cauchy, son sens. Quant aux méthodes de simplification de la procédure d'analyse, elles se basaient surtout sur cette propriété de régularité.

Il est possible aussi que pour $\varepsilon = 0$ la structure et l'ordre du système soient modifiés. Dans ce cas, les méthodes proposées au chapitre précédent ne peuvent être directement utilisées. Ces problèmes seront dits *singuliers*. Ils correspondent à l'existence de « couches frontières », une situation assez typique en analyse des systèmes.

L'analyse de tels systèmes nous confronte à une étonnante propriété des procédures d'analyse numérique. Les systèmes singuliers sont, formellement, bien plus difficiles à analyser numériquement que ceux dont les seconds membres sont des fonctions régulières variant lentement. En effet, la singularité fait croître très rapidement une partie des variables sur certains intervalles de temps. Les schémas aux différences classiques cessent d'être efficaces dans ces conditions, il faut alors passer à des approximations aux différences plus compliquées qui freinent considérablement les calculs et les rendent parfois transcendants. Mais cette transcendance permet parfois de simplifier le système. Nous verrons plus bas que la singularité exprime très souvent que le système contient des degrés de liberté parasites qui peuvent être négligés.

Dans ce chapitre, on se penchera sur l'analyse de quelques classes de cas singuliers.

La possibilité de remplacer un système d'équations par un autre, plus simple, est souvent liée à l'existence de petits paramètres en les dérivées.

Considérons le système

$$\dot{x} = X(x, y, t), \qquad \varepsilon\dot{y} = Y(x, y, t), \tag{1.1}$$

où $x$ et $y$ sont des fonctions de dimensions respectives $n$ et $m$.

Posons pour ce système le problème de Cauchy suivant:

$$t = t_0, \qquad x(t_0) = x_0, \qquad y(t_0) = y_0. \tag{1.2}$$

La résolution numérique du problème de Cauchy (1.1), (1.2) soulève quelques difficultés. En effet, lorsque $\varepsilon \to 0$, la dérivée de la fonction $y$ sera grande: $\dot{y} = O(1/\varepsilon)$. Donc, la variable $y$ est rapide et

l'on est de nouveau confronté aux difficultés signalées au chapitre précédent. Mais l'usage direct des méthodes du chapitre IV est impossible, car le système (1.1) n'est pas de la même nature que ceux du chapitre précédent. Mettons la deuxième équation (1.1) sous la forme

$$\dot{y}=\frac{1}{\varepsilon} Y(x, y, t)=f(x, y, t, \varepsilon).$$

Le second membre de cette équation n'est plus une fonction analytique du paramètre ε et le développement de la solution en une série de ε n'a plus de sens.

Néanmoins la présence d'un petit paramètre dans le système (1.1) laisse la porte ouverte à une simplification. L'idée première de l'analyste est de négliger la dérivée du vecteur $y$, puisqu'elle est multipliée par un petit facteur, donc la quantité $\varepsilon\dot{y}$ est petite. Le système (1.1) est alors remplacé par le suivant:

$$\dot{x} = X(x, y, t), \qquad Y(x, y, t) = 0. \tag{1.3}$$

Ce système sera dit *générateur*. Il est plus simple que le système primitif. D'abord, il est d'ordre $n$ et pas $n+m$. La deuxième équation du système nous permet de déterminer immédiatement la fonction vectorielle $y(t)$ comme une fonction de $x$ et de $t$:

$$y(t) = y^0(x, t). \tag{1.4}$$

En portant cette expression dans la première équation du système (1.1), on obtient

$$\dot{x} = X(x, y^0(x, t), t). \tag{1.5}$$

Le système (1.5) ne renferme que $x$, c'est-à-dire qu'il est d'ordre $n$. Son analyse numérique est bien plus aisée que celle du système primitif et pas seulement parce qu'il est d'ordre inférieur mais aussi parce que toutes ses variables sont lentes, d'où la possibilité d'effectuer l'intégration avec un grand pas.

Mais cette procédure appelle de nombreuses questions. Souvenons-nous tout d'abord qu'il faut résoudre le problème de Cauchy (1.2) pour le système d'équations (1.1), c'est-à-dire trouver la trajectoire du système (1.1) qui vérifie la condition (1.2). Le système générateur (1.3) ne nous permet pas de résoudre le problème de Cauchy, car la valeur initiale du vecteur $y$ est bien définie: $y(t_0) = y^0(x_0, t_0)$ et dans le cas général $y^0(x_0, t_0) \neq y_0$. Donc, on peut formuler pour le système (1.1) des conditions (initiales ou, par exemple, aux limites) qui n'ont pas de sens pour le système (1.3) ou (1.5).

La première question qui se pose est donc de savoir dans quel sens la solution du système (1.3) peut être proche de celle du système (1.1). Autrement dit, comment peut-on à l'aide de la solution

du système générateur (1.3) construire la solution du problème de Cauchy (1.1), (1.2) ? Cette question, classique pour toute théorie asymptotique, est particulièrement épineuse ici, car elle ne peut être résolue dans le cadre de la théorie classique du petit paramètre. En effet, lorsque $\varepsilon \to 0$, on est conduit à un système d'un autre ordre, qui est doué de propriétés totalement différentes. La solution du système (1.1) n'est plus une fonction continue du paramètre.

L'analyse du problème considéré soulève de nombreuses questions n'ayant pas d'analogue dans les théories mentionnées dans le chapitre précédent. Signalons tout d'abord que la fonction $y = y^0(x, t)$ est racine de l'équation transcendante

$$Y(x, y, t) = 0. \tag{1.6}$$

Cette équation n'étant pas linéaire, divers cas peuvent se présenter :

a) l'équation (1.6) n'admet pas de racine;

b) l'équation (1.6) possède un nombre fini de racines;

c) l'équation (1.6) présente une infinité de racines et peut notamment être une identité.

Donc le choix de la racine de l'équation (1.6) au voisinage de laquelle on construira la solution approchée est un point de théorie autonome et important.

On verra plus bas que les ingénieurs ont été confrontés aux systèmes (1.1) au siècle dernier déjà. Mais l'étude de ces systèmes ne s'est érigée en théorie autonome que tout récemment, après le travail classique de A. Tikhonov [68] à qui l'on doit le résultat fondamental de cette théorie.

REMARQUE. Les perturbations singulières ne se sont pas posées uniquement aux ingénieurs. Elles occupent une place importante dans les travaux de mathématiciens aussi notoires que Birkhoff, Tamarkine, et autres. Cependant, seul le théorème fondamental de Tikhonov a donné la clef de leur étude.

Nous ne produirons ni la formulation exacte du théorème de Tikhonov ni sa démonstration, qui est assez volumineuse, nous contentant d'en expliquer le sens.

Limitons-nous au cas où la racine $y^0(x, t)$ de l'équation (1.6) est une racine simple pour tous $x$ et $t$. Cette condition est naturelle, car dans le cas contraire il serait très difficile de donner une quelconque signification au système (1.5).

Mettons l'équation pour la fonction vectorielle $y$ du système (1.1) sous la forme

$$dy/d\tau = Y(x, y, t), \tag{1.7}$$

où $\tau = (t - t_0)/\varepsilon$. Suivant Tikhonov, on dira que le système (1.7) est *associé*. Les quantités $x$ et $t$ du système (1.7) sont traitées comme des paramètres. Dans ce cas, la racine $y^0(x, t)$ sera un point stationnaire du système (1.7) : sa solution stationnaire.

Pour deuxième condition on exigera que la solution stationnaire $y^0(x, t)$ soit asymptotiquement stable. Désignons par $y(\tau, x, t)$ la solution du système (1.7) qui satisfait la condition initiale

$$\bar{y}_0 = y(0, x, t).$$

On admettra que pour tout état initial $\bar{y}_0$ situé dans un voisinage assez petit du point $y^0(x, t)$ et pour tous $x$ et $t$ fixes, est satisfaite la *condition de stabilité asymptotique*

$$\lim_{\tau\to\infty} y(\tau, x, t) = y^0(x, t). \tag{1.8}$$

Cette condition est assez légitime, car si elle n'est pas remplie, il est peu probable que l'état stationnaire $y^0(x, t)$ puisse être utilisé pour l'approximation des solutions du système (1.7). On appellera *tikhonoviens* des systèmes (1.1) pour lesquels la solution stationnaire $y^0(x, t)$ est asymptotiquement stable.

Posons maintenant le problème de Cauchy (1.2) avec $t_0 = 0$ pour le système (1.1):

$$x(0) = x_0,\ y(0) = y_0. \tag{1.2'}$$

En passant au système générateur, nous avons négligé la deuxième condition (1.2). Il est logique que dans le cas général il n'y aura une correspondance entre les solutions du système générateur et du système (1.1), qui sera dit *perturbé*, que si la valeur initiale de la variable $y(t)$ est astreinte à certaines conditions. Une telle condition sera l'appartenance de l'état initial au domaine d'attraction de la racine $y^0(x, t)$.

Considérons le problème de Cauchy (1.2'), (1.7) pour le système associé d'équations. Les quantités $x_0$ et $t$ étant traitées comme des paramètres, on désignera la solution de ce problème par $\hat{y}(\tau, x_0, 0)$:

$$\hat{y}(0, x_0, 0) = y_0. \tag{1.9}$$

La *condition d'attraction* exprime que la solution $\hat{y}(\tau, x_0, 0)$ est proche de la racine $y^0(x_0, 0)$. Suivant Tikhonov, elle est de la forme

$$\lim_{\tau\to\infty} \hat{y}(\tau, x_0, 0) = y^0(x_0, 0). \tag{1.10}$$

Il est évident que la condition de stabilité asymptotique de la racine $y^0(x, t)$ et la condition d'attraction ne sont pas équivalentes. En tous les cas, la condition d'attraction ne résulte pas de la condition de stabilité. En effet, la stabilité asymptotique n'exprime qu'une chose: si les valeurs initiales sont assez proches de $y^0(x, t)$, alors la solution du système d'équations (1.7) tendra vers $y^0(x, t)$ pour $\tau \to 0$ quels que soient $x$ et $t$. Cela ne veut pas nécessairement dire

que la solution du système associé qui satisfait les conditions initiales (1.2′) approchera la solution du problème de Cauchy relatif au système d'équations primitif (1.1) : la valeur initiale $y_0$ peut différer d'autant que l'on veut de $y^0(x_0, 0)$.

La figure 5.1 illustre ce qui vient d'être dit. Cette figure représente les trajectoires du système associé pour $t = 0$, $x = x_0$. Si $\bar{y}_0 \in ]y_1, y_2[$, alors les trajectoires tendent vers la racine $y^0(x_0, 0)$ pour $\tau \to \infty$ (c'est-à-dire pour $\varepsilon \to 0$). Pour les valeurs assez élevées de la différence $|\bar{y}_0 - y^0(x, 0)|$, elles se conduisent n'importe comment. La condition énoncée signifie que les valeurs initiales doivent appartenir au domaine $]y_1, y_2[$ qui en l'occurrence est un domaine d'attraction. La condition d'attraction peut être interprétée d'une autre manière : cette condition restreint le choix de la racine $y^0(x, t)$ si seulement l'équation $Y(x, y, t) = 0$ admet plus d'une racine.

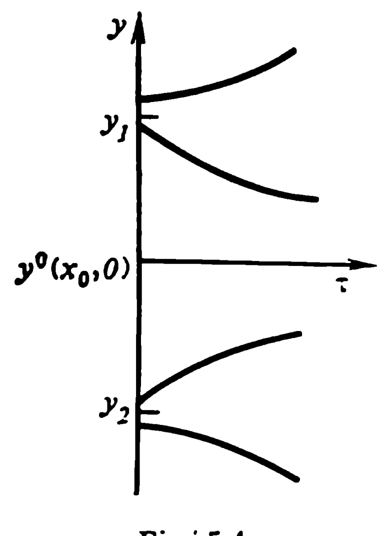


Fig. 5.1

Signalons que pour un système associé linéaire, la condition d'attraction de la racine est automatiquement réalisée si l'état stationnaire (qui est unique) est asymptotiquement stable. Illustrons ce qui vient d'être dit sur un exemple.

Supposons que la deuxième équation du système (1.1) est scalaire et est de la forme

$$\varepsilon y = -a(x, t)\, y,$$

où $a(x, t) > 0$ pour tous $x$ et $t$. Cette équation possède un seul état stationnaire $y = 0$ qui est stable, puisque la solution de l'équation associée est de la forme

$$y(\tau) = y_0 \exp\{-a(x, t)\tau\}.$$

On a $\lim\limits_{\tau\to\infty} y(\tau) = 0$ quels que soient $y_0$ et $t$. Il est évident que la condition d'attraction (pour $t = 0$) est remplie.

Outre les conditions énumérées, nous admettrons toujours que les problèmes de Cauchy relatifs aux systèmes (1.1), (1.5) et (1.7) admettent une solution et que toutes les trajectoires définies sur l'intervalle de temps fini considéré appartiennent à un domaine fini de l'espace.

Dans ces conditions, le théorème fondamental de A. Tikhonov dit que pour $\varepsilon \to 0$, la solution du problème de Cauchy relatif au système perturbé (1.1) avec les conditions initiales (1.2′) converge vers la solution du problème de Cauchy relatif au système générateur

(1.5) avec les conditions initiales

$$t = 0, \quad x(0) = x_0, \tag{1.2''}$$

les fonctions vectorielles $x(t)$ et $y(t)$ étant uniformément convergentes en $t$, la première sur tout intervalle fini $[0, T]$, la seconde sur l'intervalle $0 < \alpha \leqslant t < T$, où $\alpha$ est un nombre strictement positif aussi petit que l'on veut.

On voit qu'il n'y a pas convergence au point $t = 0$.

On aurait pu le pressentir, puisque la solution du système générateur n'est en aucune façon liée aux conditions initiales du système perturbé (1.1).

Ainsi, le théorème de Tikhonov établit les conditions qui permettent d'utiliser la solution de l'équation (1.5) pour approcher celle du système (1.1). De plus, ce théorème estime l'erreur. Si par $x(t, \varepsilon)$ et $y(t, \varepsilon)$ on désigne la solution du problème de Cauchy (1.1), (1.2′) et par $x^0(t)$ et $y^0(t)$, celle du système générateur (1.3) sous la condition (1.2″), alors le théorème nous dit que

$$\begin{aligned} x(t, \varepsilon) &= x^0(t) + O(\varepsilon) \quad \forall\, t \in [0, T], \\ y(t, \varepsilon) &= y^0(t) + O(\varepsilon) \quad \forall\, t \in [\alpha, T], \qquad \alpha > 0. \end{aligned} \tag{1.11}$$

Comment améliorer l'approximation (1.11)? Peut-on établir une approximation uniforme sur l'intervalle $[0, T]$ tout entier non seulement par rapport à $x$ mais aussi par rapport à la variable « rapide » $y$? Ces questions ont fait l'objet de nombreuses recherches, notamment par les élèves de Tikhonov. Pour répondre à ces questions, il faut faire appel à des procédures itératives spéciales ou représenter la solution par des séries spéciales. Il n'y a aucune analogie ici avec les méthodes régulières du petit paramètre. La raison étant dans les particularités du comportement de la solution au voisinage de l'origine des coordonnées. Il se trouve que la solution ne peut être représentée que par une somme

$$\begin{aligned} x &= x_1(t, \varepsilon) + x_f(t, \varepsilon), \\ y &= y_1(t, \varepsilon) + y_f(t, \varepsilon), \end{aligned} \tag{1.12}$$

où les premiers termes sont développables en séries de $\varepsilon$ et les fonctions $x_f$ et $y_f$ sont appelées *fonctions frontières*: elles compensent l'erreur des conditions aux limites et se développent suivant les puissances négatives de $\varepsilon$.

Le terme « fonction frontière », ou « fonction de couche limite » comme on l'appelle parfois, n'est pas très heureux. Les fonctions $x_f(t, \varepsilon)$ et $y_f(t, \varepsilon)$ compensent l'écart entre la solution exacte $x(t, \varepsilon)$, $y(t, \varepsilon)$ et $x^0(t)$, $y^0(t)$ au voisinage de $t = 0$, écart résultant du fait que $y^0(x^0, t) \neq y_0$. La compensation de l'erreur des conditions aux limites n'est que l'un des problèmes pour lesquels sont introduites les fonctions de couche limite.

La technique de construction des développements en séries représentant des fonctions de couche limite n'est pas triviale. Nous n'avons pas l'intention de faire un exposé tant soit peu complet du formalisme développé à ce jour (cf. [69]). Nous nous bornerons à analyser le cas où l'on aura besoin seulement des fonctions frontières obtenues par la résolution de certaines équations différentielles linéaires. Et nous nous attarderons tout d'abord sur la construction d'une solution approchée de la forme

$$x = x^0(t), \qquad y = y^0(t) + y_{f1}(t), \tag{1.13}$$

où $y_{f1}$ est le premier terme du développement de la fonction frontière $y_f(t, \varepsilon)$. La représentation (1.13) nous fournit déjà une meilleure approximation que (1.11):

$$y = y^0 + y_{f1}(t) + O(\varepsilon^2) \qquad \forall t \in [\alpha, T], \qquad \alpha > 0, \tag{1.14}$$

ou

$$y = y^0 + y_{f1}(t) + O(\varepsilon) \qquad \forall t \in [0, T] \tag{1.14'}$$

c'est-à-dire qu'on obtient une approximation uniforme à $\varepsilon$ près sur $[0, T]$.

Etudions essentiellement le cas où l'équation pour les variables rapides est linéaire en $y$:

$$\varepsilon \frac{dy}{dt} = A(x, t)\, y + b(x, t, \varepsilon). \tag{1.15}$$

Ce cas présente de l'intérêt pour les applications. En effet, considérons de nouveau le système (1.1) et supposons que les conditions du théorème de Tikhonov sont réunies. Alors la solution $x^0(t)$, $y^0(t)$ du système générateur nous fournit (si $\varepsilon$ est assez petit) une approximation satisfaisante de la solution exacte. L'étude de cette solution qui est bien plus aisée que celle du problème primitif constitue la première étape de l'analyse. Il faut ensuite préciser cette solution. Si elle est « assez bonne », alors on posera naturellement $y = y^0 + y_1$ et on linéarisera l'équation pour $y$ par rapport à $y_1$:

$$\varepsilon\left(\frac{dy^0}{dt} + \frac{dy_1}{dt}\right) = Y(x, y^0 + y_1, t) = A_1(x, y^0, t)\, y_1 + b_1, \tag{1.16}$$

où $A_1 = \left(\frac{\partial Y}{\partial y}\right)_{y=y^0}$. Le développement (1.16) tient compte du fait que $y^0$ est une racine de l'équation $Y(x, y^0, t) = 0$. Comme $dy^0/dt$ est une fonction connue du temps, on obtient en définitive une équation de la forme (1.15).

Ainsi, nous porterons notre attention essentiellement sur l'analyse d'un système de la forme

$$\dot{x} = X(x, y, t), \qquad \varepsilon\dot{y} = A(x, t)\, y + b(x, t, \varepsilon). \tag{1.17}$$

Ce système est beaucoup plus simple que le système primitif (1.1). Bâtissons une théorie des perturbations pour l'analyser.

Tout d'abord, remplaçons la deuxième équation de ce système par la suivante:

$$\varepsilon \frac{dy}{dt} = A(x^0,\ t)\, y + b(x^0,\ t,\ \varepsilon), \tag{1.18}$$

où $x^0(t)$ est la solution du problème (1.5), (1.2″).

Linéarisons ensuite la première équation du système (1.17) par rapport à $y$:

$$\dot{x} = X_1(x,\ t) + X_2(x,\ t)\, y. \tag{1.19}$$

Enfin, posons $x = x^0 + z$ et linéarisons l'équation (1.19) par rapport à $z$:

$$\dot{z} = B(t)\, z + X_2(x^0,\ t)\, y. \tag{1.20}$$

Ces méthodes de la théorie des perturbations sont largement appliquées en technique et leur signification mathématique est claire. Le premier schéma d'analyse du genre est probablement l'œuvre de de Sparre, un capitaine de l'artillerie française qui le premier étudia les trajectoires de projectiles tournants. Ses travaux qui datent des années 70 du siècle dernier ne faisaient intervenir aucune construction mathématique et les méthodes d'analyse préconisées furent de nouveau découvertes au XX<sup>e</sup> siècle par Birkhoff, Tamarkine et d'autres.

Les schémas de calcul proposés par de Sparre continuèrent de faire recette en artillerie malgré l'apparition de profondes investigations mathématiques des systèmes de la forme (1.17). Mais l'agrégation de toutes ces recherches mathématiques et techniques n'eut lieu qu'aux années 30 avec les travaux de D. Ventsel et V. Pougatchev, professeurs à l'Académie militaire Joukovski (cf. [60]).

REMARQUE. La théorie des systèmes tikhonoviens est largement appliquée dans les problèmes d'optimisation. Nous reviendrons sur cette question au chapitre VIII. Signalons ici une circonstance importante pour les applications.

La communication [33] a mis en évidence le lien étroit qui unit la théorie des perturbations singulières et la recherche du minmax

$$J = \min_{x} \max_{y} F(x,\ y). \tag{1.21}$$

Nous avons déjà examiné des problèmes de cette nature au § 4 du chap. I dans le cadre de la prise de décision en présence d'indéterminations.

Rappelons que le problème (1.21) se décompose en un problème intérieur de recherche de

$$\varphi(x) = \max_{y} F(x,\ y) = F(x,\ y(x)) \tag{1.22}$$

et un problème extérieur de minimisation de la fonction $\varphi(x)$:

$$J = \min_x \varphi(x).$$

Supposons que la fonction $F(x, y)$ est deux fois dérivable par rapport à ses deux arguments et que le problème intérieur (1.22) admet une solution unique pour tout $y$. Alors, une condition nécessaire de maximum de (1.22) est

$$\frac{\partial F(x, y(x))}{\partial y} = 0. \qquad (1.23)$$

Si l'on admet que la matrice $F_{yy}(x, y)$ est partout définie négative, alors la fonction $y(x)$ est dérivable et est solution de l'équation

$$F_{yy}(x, y(x)) \frac{dy}{dx} + F_{yx}(x, y(x)) = 0.$$

Pour que la fonction $\varphi(x)$ présente un minimum en $x = x_*$, il est nécessaire que, premièrement, le point $x_*$ soit stationnaire, c'est-à-dire en vertu de (1.23) que

$$\frac{d\varphi}{dx} = F_x(x_*, y_*) = 0, \quad y_* = y(x_*),$$

et, deuxièmement, que la matrice

$$N(x_*, y_*) = F_{xx}(x_*, y_*) - F_{xy}(x_*, y_*) F_{yy}^{-1}(x_*, y_*) F_{yx}(x_*, y_*),$$

soit semi-définie positive, c'est-à-dire que

$$N(x_*, y_*) \geqslant 0.$$

On est conduit à la conclusion suivante: pour que le point $(x_*, y_*)$ soit solution du problème (1.21) (c'est-à-dire un point de minimum local) il est suffisant qu'il soit stationnaire, c'est-à-dire que

$$F_x(x_*, y_*) = F_y(x_*, y_*) = 0, \qquad (1.24)$$

et que les matrices $F_{yy}(x_*, y_*)$ et $N(x_*, y_*)$ soient définies positives. Supposons qu'un tel point $(x_*, y_*)$ existe. Question: quelle méthode numérique utiliser pour le trouver? Le travail [33] propose de chercher les points limites (pour $t \to \infty$) des solutions du problème de Cauchy

$$\begin{aligned} \frac{dx}{dt} &= -F_x(x, y), \quad x(0) = x_0, \\ \varepsilon \frac{dy}{dt} &= F_y(x, y), \quad y(0) = y_0, \end{aligned} \qquad (1.25)$$

où $\varepsilon \ll 1$ est un petit paramètre. Il est immédiat de voir que ce système est un cas particulier du système (1.1). L'analogue de l'équation (1.6) est ici le problème (1.22) qui se ramène à la résolution de l'équation (1.23). On a la proposition suivante:

THÉORÈME. *Si la fonction $F(x, y)$ est bicontinûment dérivable au voisinage du point stationnaire $(x_*, y_*)$ et les matrices $F_{yy}(x, y)$ et $N(x, y)$, définies positives, alors il existe un nombre $\bar{\varepsilon}$ tel que pour tout $\varepsilon \in ]0, \bar{\varepsilon}[$ les solutions du système* (1.25) *convergent localement vers le point $(x_*, y_*)$ lorsque $t \to \infty$.*

Si dans (1.25) on admet que $\varepsilon$ est un grand paramètre ($\varepsilon \gg 1$) et si les conditions suffisantes de maximin local sont réunies, alors les solutions du système (1.25) convergent pour $t \to \infty$ vers des points qui son solutions du problème de maximin:

$$\max_{\{y\}} \min_{\{x\}} F(x, y).$$

Si la fonction $F(x, y)$ est strictement convexe-concave, alors on peut poser $\varepsilon = 1$ dans (1.25), et l'on est conduit à la méthode numérique d'Arrow-Hurwitz de recherche des points cols.

La signification de ce théorème est évidente. En effet, composons l'équation différentielle de la courbe gradient pour déterminer le minimum de la fonction $\varphi(x)$. Cette équation peut s'écrire

$$\frac{dx}{dt} = -\frac{dF}{dx}\Big| = -F_x(x, y), \tag{1.26}$$

où $y$ est une fonction de $x$ définie par la condition

$$F(x, y) \Rightarrow \max_y.$$

La fonction $F(x, y)$ étant dérivable par hypothèse, on cherchera $y$ à partir de l'équation

$$F_y(x, y) = 0. \tag{1.26'}$$

Si les seconds membres du système (1.25) satisfont les conditions du théorème de Tikhonov, alors sa solution tend vers celle du système (1.26), (1.26') lorsque $\varepsilon \to 0$. La validité des conditions de Tikhonov est assurée par la définition positive des matrices $F_{yy}$ et $N$.

Cette approche est largement utilisée dans la résolution des problèmes de programmation non linéaire et en théorie des jeux.

## § 2. Problème linéaire

Le théorème fondamental de Tikhonov examiné au paragraphe précédent ouvre de très intéressantes et importantes perspectives pour l'étude des systèmes complexes, la réduction de leur ordre et l'élaboration d'une théorie des perturbations, ainsi que pour la construction de méthodes simplifiées et économiques de calcul numérique. Ces méthodes permettent de réduire de plusieurs fois le temps d'occupation de l'ordinateur et possèdent un vaste spectre d'application, puisque le cas où les dérivées d'une partie des coordonnées de

phase sont multipliées par un petit paramètre est classique. Soit donc un système réductible à la forme

$$\dot{x} = X(x, y, t), \qquad \varepsilon\dot{y} = Y(x, y, t), \tag{2.1}$$

où $x$ et $y$ sont des vecteurs de dimension respective $n$ et $m$.

Si les conditions du théorème de Tikhonov sont remplies, alors l'approximation d'ordre zéro

$$\dot{x}^0 = X(x^0, y^0, t), \quad Y(x^0, y^0, t) = 0 \tag{2.2}$$

peut être prise pour solution approchée avec une erreur partout de l'ordre de $O(\varepsilon)$ sauf à l'origine. Le système (2.2) ne renferme pas de termes à variation rapide, puisque le vecteur $y$, solution de l'équation

$$\dot{y} = \frac{1}{\varepsilon} Y(x, y, t),$$

est remplacé par la racine de la deuxième équation du système (2.2) :

$$y = y^0(x, t). \tag{2.3}$$

Des méthodes ont été développées (cf. [69]) qui permettent de construire sur la base de la solution du système (2.2) des séries asymptotiques destinées à approcher uniformément la solution du système (2.1) vérifiant les conditions initiales

$$t = 0, \qquad x(0) = x_0, \qquad y(0) = y_0 \tag{2.4}$$

sur tout intervalle fini avec n'importe quelle précision. Mais pour analyser les systèmes écologiques, techniques, économiques, on peut souvent se contenter des théories des perturbations mentionnées à la fin du § 1. Nous considérons ici certaines variantes de la théorie des perturbations, nécessitant l'analyse de l'équation

$$\varepsilon\dot{y} = A(t)\, y + b(t, \varepsilon). \tag{2.5}$$

Au § 1 nous avons parlé des équations de la forme (2.5) lors de la linéarisation de la deuxième équation du système (2.1) au voisinage du point $y = y^0(x, t)$ sous réserve que les valeurs intiales $y^0(x_0, 0)$ et $y_0$ soient voisines. Mais les équations de la forme (2.5) présentent un intérêt en soi. En posant $\lambda = 1/\varepsilon$, on peut mettre l'équation (2.5) sous la forme

$$\dot{y} = \lambda A(t)\, y + \lambda b(t, 1/\lambda). \tag{2.5'}$$

Sous cette forme l'équation (2.5) se rencontre dans de nombreux problèmes d'application.

a) *Systèmes homogènes. Cas de racines simples.* Considérons des équations de la forme

$$\dot{y} + \lambda A(t)\, y = 0 \tag{2.6}$$

où $\lambda = 1/\varepsilon$, et étudions le comportement des solutions pour $\lambda \to \infty$ *).

Cherchons les solutions particulières du système (2.6) sous la forme

$$y = \exp\left\{\int_0^t \lambda\mu(t)\,dt\right\} z(\lambda, t), \tag{2.7}$$

où $\mu(t)$ est une racine de l'équation caractéristique

$$|A + \mu E| = 0. \tag{2.8}$$

Les racines de l'équation (2.8) seront des fonctions du temps. Nous n'étudierons que le seul cas où l'équation (2.8) admet des racines $\mu_i(t)$ distinctes sur l'intervalle $[0, T]$ et les fonctions $\mu_i(t)$ ne s'annulent en aucun point de $[0, T]$.

Cherchons la fonction inconnue $z(\lambda, t)$ sous forme de la série

$$z(\lambda, t) = z_0(t) + \lambda^{-1} z_1(t) + \ldots \tag{2.9}$$

On se propose d'indiquer une procédure qui à chaque racine de l'équation (2.8) associe une expression (asymptotique) approchée de la solution particulière du système (2.6).

En portant (2.7) et (2.9) dans l'équation (2.6) et en identifiant les coefficients des mêmes puissances de $\lambda$, on obtient les équations suivantes

$$(A + \mu E)\, z_0 = 0, \tag{2.10}$$

$$(A + \mu E)\, z_1 = -\dot{z}_0, \tag{2.11}$$

etc.

Comme $\mu(t)$ est une racine de l'équation (2.8), le système (2.10) admet une solution et les composantes du vecteur $z_0$ peuvent être déterminées à un facteur multiplicatif près. Cette proposition revient à dire qu'une composante du vecteur $z_0$ (par exemple $z_0^1$) reste indéterminée et les autres peuvent être exprimées en fonction d'elle.

Mettons le système (2.10) sous la forme suivante:

$$\begin{aligned}
(a_{22} + \mu)\, z_0^2 + a_{23} z_0^3 + \ldots &= -a_{21} z_0^1, \\
a_{32} z_0^2 + (a_{33} + \mu)\, z_0^3 + \ldots &= -a_{31} z_0^1, \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots& \\
a_{n2} z_0^2 + \ldots + (a_{nn} + \mu)\, z_0^n &= -a_{n1} z_0^1.
\end{aligned} \tag{2.12}$$

*) Le système traité ici est un cas particulier des systèmes $\dot{y} = \lambda^k A(t, \lambda)\, y$ étudiés en détails par Ya. Tamarkine [65].

Désignons par $\Delta_{1i}$ les cofacteurs des éléments de la première ligne du déterminant $|A + \mu E|$. La solution du système (2.12) peut alors s'écrire

$$z_0^k = -\frac{\Delta_{1k}}{\Delta_{11}} z_0^1, \quad k = 2,\ 3,\ \ldots,\ n. \tag{2.13}$$

Le rang de la matrice $A + \mu E$ sera égal à $n - 1$, puisque nous avons admis que toutes les racines de l'équation (2.8) étaient simples. Donc l'un au moins des mineurs d'ordre $n - 1$, pour fixer les idées $\Delta_{11}$, sera non nul.

Considérons maintenant le système (2.11). Son déterminant est nul de par le choix de $\mu$, et sa matrice est de rang $n - 1$. Donc pour que ce système admette une solution, il est nécessaire et suffisant que le rang de la matrice élargie (c'est-à-dire la matrice obtenue en adjoignant la colonne des seconds membres à la matrice du système) soit aussi égal à $n - 1$. Le rang de la matrice $A + \mu E$ étant lui aussi égal à $n - 1$, il existe entre les éléments de ses lignes une relation linéaire que l'on peut écrire comme suit si l'on développe le déterminant $|A + \mu E|$ suivant les éléments de la première colonne

$$a_{11} + \mu = c_2 a_{21} + c_3 a_{31} + \ldots + c_n n_{n1}, \tag{2.14}$$

où

$$c_j = -\Delta_{j1}/\Delta_{11}. \tag{2.15}$$

Donc pour que le système (2.11) admette une solution, il est nécessaire et suffisant qu'entre les éléments de la colonne du second membre on ait la même relation (2.14). Cela signifie que

$$\dot z_0^1 = c_2 \dot z_0^2 + c_3 \dot z_0^3 + \ldots + c_n \dot z_0^n. \tag{2.16}$$

Remplaçons $z_0^k$ par son expression (2.13)

$$\dot z_0^k = -\frac{\Delta_{1k}}{\Delta_{11}} \dot z_0^1 - z_0^1 \frac{d}{dt}\left(\frac{\Delta_{1k}}{\Delta_{11}}\right). \tag{2.17}$$

L'égalité (2.16) devient en définitive:

$$-\dot z_0^1 = \left(c_2 \frac{\Delta_{12}}{\Delta_{11}} + c_3 \frac{\Delta_{13}}{\Delta_{11}} + \ldots + c_n \frac{\Delta_{1n}}{\Delta_{11}}\right) \dot z_0^1 +$$
$$+ \left\{c_2 \frac{d}{dt}\left(\frac{\Delta_{12}}{\Delta_{11}}\right) + c_3 \frac{d}{dt}\left(\frac{\Delta_{13}}{\Delta_{11}}\right) + \ldots + c_n \frac{d}{dt}\left(\frac{\Delta_{1n}}{\Delta_{11}}\right)\right\} z_0^1. \tag{2.18}$$

L'équation (2.18) est une équation différentielle ordinaire linéaire du premier ordre par rapport à $z_0^1(t)$. On peut la mettre sous la forme $\dot z_0^1 U(t) = V(t)\, z_0^1$, d'où

$$z_0^1(t) = c \exp\left\{\int_0^t \frac{V}{U}\, dt\right\}, \tag{2.19}$$

où $c$ est une constante arbitraire.

Les autres termes du développement (2.9) se déterminent suivant le même scénario : à chaque pas, il faudra résoudre une équation du premier ordre, ce qui fera apparaître une nouvelle constante arbitraire. En se donnant ces constantes, on fixe l'état initial du vecteur qui définit la solution particulière correspondant à la valeur choisie de la racine de l'équation caractéristique.

Ces raisonnements étant valables pour toutes les racines de l'équation caractéristique (2.8) et ces racines étant toutes distinctes par hypothèse, la procédure préconisée permet de construire le système complet des solutions linéairement indépendantes du système (2.6) et toutes ces solutions peuvent être exprimées par des quadratures.

Récapitulons. Supposons que le système envisagé est de la forme (2.1) et que les conditions du théorème de Tikhonov sont réunies. On commence alors par construire la solution du système (2.2). Celle-ci nous donne une solution approchée à $O(\varepsilon)$ près. Mais ce n'est pas la solution qu'il nous faut, car le problème de Cauchy n'a été résolu que pour la variable $x$. La variable $y$ s'obtient par la résolution de l'équation transcendante

$$Y(x, y, t) = 0. \tag{2.20}$$

Nous avons désigné cette solution par $x^0(t)$, $y^0(t)$. Cette solution ne réalise pas la deuxième condition initiale (2.4) :

$$t = 0, \qquad y(0) = y_0. \tag{2.20'}$$

Pour satisfaire la condition (2.20′), on construit la fonction frontière $y_f$. Cette fonction est solution d'une équation linéaire de la forme (2.5). Nous avons indiqué une méthode de construction de l'intégrale générale de cette équation dans le cas où $b(t) \equiv 0$. En portant cette intégrale dans (2.20′), on peut déterminer les constantes arbitraires et achever la construction de la fonction frontière. La solution est alors de la forme

$$x = x^0(t), \qquad y = y^0(t) + y_f(t). \tag{2.21}$$

La première expression approche la solution exacte à $O(\varepsilon)$ près uniformément sur $[0, T]$. Grâce au terme $y_f$ ajouté, la fonction $y(t)$ est douée de la même propriété et réalise la condition (2.20′).

REMARQUE. A strictement parler, les raisonnements effectués ne sont valables que si toutes les racines de l'équation caractéristique $|A + \mu E| = 0$ ont des parties réelles strictement négatives, car dans ce cas seul la stabilité asymptotique a lieu et le théorème de Tikhonov est valable, donc les formules (2.21) fournissent l'approximation nécessaire. Mais les résultats acquis dans ce numéro peuvent recevoir une interprétation qui sort du cadre du théorème de Tihkonov et de l'analyse du système (2.1). L'équation (2.6) est un objet d'étude autonome et important. La procédure développée fournit une méthode de construction d'estimations asymptotiques pour cette équation. Il s'avère qu'on peut utiliser à cet effet toute tranche finie de la série (2.9). Posons

$$z^{(k)}(t, \lambda) = z_0(t) + \lambda^{-1} z_1(t) + \ldots + \lambda^{-k} z_k(t).$$

Alors la fonction $y^{(k)} = \exp\left\{\lambda \int\limits_0^t \mu dt\right\} z^{(k)}(t, \lambda)$ approche la solution particulière correspondante à $O(1/\lambda^{k+1})$ près lorsque $\lambda \to \infty$.

Voyons maintenant quelques cas particuliers.

b) *Cas d'une seule équation du premier ordre.* Considérons l'exemple le plus élémentaire où le système qui définit les fonctions frontières se réduit à une équation scalaire du premier ordre. A ce cas correspond par exemple le système

$$\dot{x} = X(x, y, t), \qquad \varepsilon \dot{y} = -a(x)\, y. \tag{2.22}$$

Le système générateur sera

$$\dot{x}^0 = X(x^0, 0, t),\ y^0 \equiv 0 \tag{2.23}$$

avec la condition

$$x(0) = x_0. \tag{2.24}$$

La résolution numérique du problème de Cauchy (2.23), (2.24) nous donne $x^0 = x^0(t)$. En portant $x^0$ dans la deuxième équation (2.22), on obtient une équation qui doit être satisfaite par la fonction frontière $y_f$:

$$\dot{y}_f = -\lambda a(x^0(t))\, y_f = -\lambda \hat{a}(t)\, y_f, \qquad \lambda = 1/\varepsilon. \tag{2.25}$$

Il n'est pas indispensable de recourir ici aux approximations asymptotiques, puisque l'équation (2.25) admet une solution exacte

$$y_f = C \exp\left\{-\lambda \int\limits_0^t \hat{a}(\tau)\, d\tau\right\},$$

où $C = y_0 - y^0(0) = y_0$. La solution approchée du système (2.22) vérifiant la condition initiale

$$x(0) = x_0, \qquad y(0) = y_0$$

est de la forme

$$x = x^0(t), \quad y = y_f = y_0 \exp\left\{-\lambda \int\limits_0^t \hat{a}(\tau)\, d\tau\right\}.$$

La solution exacte du problème initial de Cauchy s'écrit

$$x = x^0 + O(1/\lambda), \qquad y = y_f + O(1/\lambda).$$

et cette expression approchera la solution exacte uniformément sur l'intervalle $[0, T]$ tout entier.

Si l'on veut calculer la composante lente $x(t)$ avec une précision plus élevée, il faut étudier encore un système perturbé:

$$\dot{x} = X(x, y^0(x^0, t) + y_f, t), \qquad \dot{y}_f = -\lambda \hat{a}(t)\, y_f. \tag{2.26}$$

L'intégration directe du système (2.26) pose des problèmes évidents, puisque la dérivée de la fonction $y_f$ est élevée. Posons donc $x = x^0 + z$ et linéarisons le système par rapport à $z$ et $y_f$. En tenant compte de l'expression de $y_f$, on obtient

$$\dot{z} = Lz + CM \exp\left\{-\lambda \int_0^t \hat{a}(\tau)\, d\tau\right\}, \tag{2.27}$$

où $L = \left(\frac{\partial X}{\partial x}\right)_{x=x^0}$ est une matrice, $M = \left(\frac{\partial X}{\partial y}\right)_{y=y^0}$ un vecteur, $C = y_0$.

Intégrons le système (2.27) par la méthode de variation des constantes arbitraires. Désignons par $Z(t)$ la matrice des solutions fondamentales du système homogène

$$\dot{z} = Lz. \tag{2.28}$$

On remarquera que l'intégration du système (2.28) ne soulève pas de difficulté, puisque le second membre de (2.28) ne contient pas de fonctions variant rapidement. Exprimons la solution de l'équation (2.27) par des quadratures en tenant compte de ce que $z(0) = 0$:

$$z(t) = Z(t) \int_0^t Z^{-1}(\tau)\, M(\tau)\, C \exp\left\{-\lambda \int_0^\tau \hat{a}(s)\, ds\right\} d\tau. \tag{2.29}$$

Le calcul de l'intégrale (2.29) peut se heurter aux difficultés classiques, mais on peut éviter ce calcul sans perdre en précision. En effet, une intégration de (2.29) par parties nous donne

$$z(t) = Z(t)\left\{-\frac{C}{\lambda\hat{a}}\left[\exp\left\{-\lambda \int_0^t \hat{a}\, ds\right\} Z^{-1}(t)\, M(t) - Z^{-1}(0)\, M(0)\right] + \right.$$
$$\left. + \frac{C}{\lambda} \int_0^t \frac{1}{\hat{a}} \exp\left\{-\lambda \int_0^\tau \hat{a}\, ds\right\} \frac{d}{d\tau}\left(Z^{-1}(\tau)\, M(\tau)\right) d\tau\right\}.$$

Si l'on intègre par parties le second terme de l'accolade extérieure, on remarque sans peine qu'il est de l'ordre de $O(1/\lambda^2)$. Donc

$$z(t) = Z(t)\left(-\frac{C}{\lambda\hat{a}}\right)\left[\exp\left\{-\lambda \int_0^t \hat{a}\, ds\right\} Z^{-1}(t)\, M(t) - Z^{-1}(0)\, M(0)\right] + $$
$$+ O\left(\frac{1}{\lambda^2}\right), \tag{2.30}$$

c'est-à-dire qu'on obtient $z(t)$ explicitement à $O(1/\lambda^2)$ près. On voit que l'expression de $z(t)$ ne contient plus d'intégrales des fonctions variant rapidement.

Donc, en représentant la solution approchée $x^*(t)$ par la somme

$$x^* = x^0 + z,$$

où $z$ est définie par la formule (2.30), on obtient une approximation uniforme de l'ordre de $1/\lambda^2$:

$$x = x^0 + O(1/\lambda^2).$$

c) *Cas d'un système de deux équations du premier ordre.* Considérons maintenant le système

$$\dot{y}_1 = \lambda [a_{11}y_1 + a_{12}y_2], \qquad \dot{y}_2 = \lambda [a_{21}y_1 + a_{22}y_2]. \qquad (2.31)$$

La technique d'analyse du système général (2.6) s'applique intégralement au système (2.31). Pour indiquer quelques traits spécifiques du résultat final, on se bornera au cas élémentaire où

$$a_{12} = -\omega^2(t), \qquad a_{11} = a_{22} = 0, \qquad a_{21} = 1,$$

c'est-à-dire au système

$$\dot{y}_1 = -\lambda\omega^2 y_2, \qquad \dot{y}_2 = \lambda y_1. \qquad (2.32)$$

Le système (2.32) est visiblement équivalent à l'équation du second ordre

$$\ddot{y} + \lambda^2\omega^2 y = 0. \qquad (2.33)$$

Cherchons la solution du système (2.32) sous la forme

$$y_1 = \exp\left\{\lambda \int_0^t \mu\, dt\right\}\left(z_1^{(0)} + \frac{1}{\lambda} z_1^{(1)} + \dots\right),$$

$$y_2 = \exp\left\{\lambda \int_0^t \mu\, dt\right\}\left(z_2^{(0)} + \frac{1}{\lambda} z_2^{(1)} + \dots\right),$$

où $\mu$ est une racine de l'équation caractéristique. Ici $\mu_1 = i\omega$, $\mu_2 = -i\omega$.

Supposons pour fixer les idées que $\mu = \mu_1 = i\omega$ et effectuons les calculs en détail. Les fonctions $z_1^{(0)}$ et $z_2^{(0)}$ satisferont les équations

$$i\omega z_1^{(0)} + \omega^2 z_2^{(0)} = 0, \qquad -z_1^{(0)} + i\omega z_2^{(0)} = 0.$$

Exprimons une fonction inconnue par l'intermédiaire de l'autre en utilisant, par exemple, la deuxième équation:

$$z_1^{(0)} = i\omega z_2^{(0)}. \qquad (2.34)$$

Pour déterminer les fonctions $z_1^{(0)}$ et $z_2^{(0)}$ nous devons, en vertu du schéma général, considérer les équations de la deuxième approxima-

tion, c'est-à-dire

$$i\omega z_1^{(1)} + \omega^2 z_2^{(1)} = -\dot{z}_1^{(0)}, \qquad -z_1^{(1)} + i\omega z_2^{(1)} = -\dot{z}_2^{(0)}.$$

Pour que ce système admette une solution, il est nécessaire et suffisant que le rang de la matrice élargie soit nul, c'est-à-dire que

$$\begin{vmatrix} i\omega & \dot{z}_1^{(0)} \\ -1 & \dot{z}_2^{(0)} \end{vmatrix} = 0,$$

d'où

$$i\omega \dot{z}_2^{(0)} = -\dot{z}_1^{(0)}. \tag{2.35}$$

Une dérivation de (2.34) nous donne

$$\dot{z}_1^{(0)} = i\dot{\omega} z_2^0 + i\omega \dot{z}_2^{(0)}.$$

En portant cette expression dans (2.35), on obtient l'équation suivante en $z_2^{(0)}$:

$$i\omega \dot{z}_2^{(0)} = -i\dot{\omega} z_2^{(0)} - i\omega \dot{z}_2^{(0)},$$

d'où

$$\dot{z}_2^{(0)} = -\frac{\dot{\omega}}{2\omega} z_2^{(0)},$$

ou

$$z_2^{(0)} = C \exp\left\{-\int_0^t \frac{\dot{\omega}}{2\omega}\, dt\right\} = C \exp\left\{-\ln \sqrt{\omega}\right\} = \frac{C}{\sqrt{\omega}},$$

et en définitive

$$z_2^{(0)} = C/\sqrt{\omega}.$$

En première approximation donc les solutions particulières du système (2.32) seront de la forme

$$y_{1,2} = \frac{C_{1,2}}{\sqrt{\omega}} \exp\left\{\pm i\lambda \int_0^t \omega\, dt\right\}.$$

En passant aux expressions réelles, on trouve

$$\begin{aligned} y_1 &= \frac{A}{\sqrt{\omega}} \cos\left(\lambda \int_0^t \omega(t)\, dt\right), \\ y_2 &= \frac{B}{\sqrt{\omega}} \sin\left(\lambda \int_0^t \omega(t)\, dt\right). \end{aligned} \tag{2.36}$$

Il est intéressant de signaler le lien unissant les méthodes exposées à la méthode WBKJ (Wentzel, Brillouin, Kramers, Jeffreys) utilisée par les ingénieurs pour le calcul des mouvements vibratoires. La méthode WBKJ a été proposée au siècle dernier déjà. Soit l'équation

$$\ddot{y} + \omega^2 y = 0. \tag{2.37}$$

Si la fonction $\omega$ était une constante, la solution de l'équation (2.37) serait de la forme $y_{1,\,2} = C_{1,\,2} \exp\{\pm i\omega t\}$. Si $\omega$ est assez grande et dépend du temps, la solution devrait être bien décrite par la fonction

$$\exp\left\{\pm i \int_0^t \omega \, dt\right\}.$$

En effet si $\omega$ est assez grande, les solutions de l'équation (2.37) oscillent rapidement et la fonction $\omega\,(t)$ ne peut pas varier rapidement au cours d'une période. Donc, à chaque instant $t$ les solutions linéairement indépendantes peuvent être approchées par les fonctions

$$y_{1\,\,2} = C_{1,\,2} \exp\{\pm i\overline{\omega\,(t)}\,t\},$$

où $\overline{\omega\,(t)}$ est la valeur moyenne de $\omega\,(t)$ sur une demi-période. A la lumière de ces suggestions, on cherchera les solutions de l'équation (2.37) sous la forme

$$y_{1.2} = \exp\left\{\pm i \int_0^t \omega \, dt\right\} z\,(t). \tag{2.38}$$

Le facteur exponentiel de (2.38) décrit une oscillation rapide. Il y a donc de fortes chances pour que la fonction $z\,(t)$ varie lentement.

En dérivant deux fois (2.38) et en portant dans (2.37), on obtient l'équation suivante par rapport à $z$:

$$\ddot{z} \pm 2i\omega\dot{z} \pm i\dot{\omega}z = 0. \tag{2.39}$$

Le terme en $\ddot{z}$ ne contient pas $\omega$; comme il est fort probable que $z\,(t)$ soit à variation lente, le terme $\ddot{z}$ peut être négligé dans (2.39). L'équation (2.39) se transforme alors en une équation du premier ordre qui s'intègre facilement:

$$z = c/\sqrt{\omega\,(t)}, \quad c = \text{const.}$$

On obtient ainsi l'approximation suivante pour les solutions linéairement indépendantes de l'équation (2.37):

$$y_{1.2} = \frac{C_{1.2}}{\sqrt{\omega\,(t)}} \exp\left\{\pm i \int_0^t \omega\,(t)\, dt\right\}.$$

Ceci n'est autre que la célèbre formule de la méthode WBKJ. En passant aux quantités trigonométriques, on obtient les formules (2.36) aux notations près.

Ces raisonnements montrent que dans l'établissement des formules (2.36), ce n'est pas le paramètre $\lambda$ qui est le plus important, mais le rapport de $\lambda$ et $\omega$.

Il est intéressant de remarquer que les formules (2.36) peuvent être acquises par des raisonnements relevant de la méthode de moyennisation.

d) *Equation avec second membre.* Jusqu'ici nous avons examiné le cas où la fonction frontière était solution de l'équation homogène. Considérons maintenant le cas général de l'équation (2.5) où $b(t) \not\equiv \not\equiv 0$. Mettons cette équation sous la forme

$$\dot{y} = \lambda A(t)\, y + \lambda b(t,\, 1/\lambda), \qquad \lambda = 1/\varepsilon. \tag{2.40}$$

Pour trouver la solution générale de l'équation (2.40), il faut d'abord construire une expression asymptotique d'une solution particulière. Cette solution, on peut la chercher par la méthode de variation des constantes arbitraires et ensuite éliminer les termes de précision redondante par une intégration par parties. Mais on peut la trouver directement sous forme d'une série suivant les puissances négatives du paramètre $\lambda$:

$$\tilde{y}(t,\, \lambda) = \tilde{y}^{(0)}(t) + \lambda^{-1}\tilde{y}^{(1)} + \lambda^{-2}\tilde{y}^{(2)} + \ldots \tag{2.41}$$

En portant la série (2.41) dans l'équation (2.40) et en identifiant les coefficients des mêmes puissances de $\lambda$, on obtient les expressions suivantes pour la détermination des termes de la série (2.41):

$$\tilde{y}^{(0)} = -A^{-1}b,\ \tilde{y}^{(1)} = -A^{-1}\frac{d\tilde{y}^{(0)}}{dt},\ \ldots,\quad \tilde{y}^{(k)} = -A^{-1}\frac{d\tilde{y}^{(k-1)}}{dt},\ \ldots \tag{2.42}$$

Les formules (2.42) permettent de calculer successivement n'importe quel nombre de termes de la série.

La procédure développée ne revêt pas un caractère formel: toute tranche finie de la série (2.41) nous donne une approximation asymptotique d'une solution particulière de l'équation (2.40), c'est-à-dire que

$$\begin{gathered} \tilde{y}(t,\, \lambda) = \tilde{y}^{(0)}(t) + O\left(\frac{1}{\lambda}\right), \\ \tilde{y}(t,\, \lambda) = \tilde{y}^{(0)}(t) + \frac{1}{\lambda}\tilde{y}^{(1)}(t) + O\left(\frac{1}{\lambda^2}\right), \\ \ldots\ldots\ldots\ldots\ldots\ldots \end{gathered} \tag{2.43}$$

Ainsi, la procédure exposée permet d'approcher une solution particulière avec n'importe quelle précision sur tout intervalle de

temps fini donné. Le calcul des termes de la série (2.41) peut être effectué par des formules finies *).

Les numéros précédents de ce paragraphe ont été consacrés aux méthodes de construction des expressions asymptotiques de l'intégrale générale de l'équation homogène associée à (2.40). Désignons ces solutions par $y_i$ $(t, \lambda)$. L'expression générale de la fonction frontière peut alors s'écrire

$$y(t, \lambda) = \sum_i c_i y_i(t, \lambda) + \tilde{y}(t, \lambda), \tag{2.44}$$

où les fonctions $y_i$ et $\tilde{y}$ sont calculées avec la même précision par rapport à $\lambda$. Les constantes $c_i$ se déduisent de la condition

$$t = 0, \qquad y(0) = y_0.$$

On vient de montrer comment construire des fonctions frontières pour des équations linéaires générales.

Remarque importante: on a vu que la théorie développée permettait d'éviter l'intégration des fonctions variant rapidement et d'exprimer les fonctions qui nous intéressaient par des quadratures simples. Elle permet d'autre part d'estimer la précision. Mais ces estimations ne sont pas toujours suffisantes en pratique. En effet, que signifie, par exemple, l'égalité

$$x = x^0 + O(\varepsilon^h)? \tag{2.45}$$

Elle exprime qu'à tout instant $t$ appartenant à un intervalle de temps fini, on a la majoration

$$|x - x^0| \leqslant C\varepsilon^k.$$

On peut donc seulement garantir l'existence d'une constante $C$ telle que pour tout $t \in [0, T]$ la fonction $C\varepsilon^k$ majore la valeur absolue $|x - x^0|$.

Donc, l'estimation (2.45) ne concerne que la vitesse avec laquelle $x - x^0$ tend vers 0 lorsque $\varepsilon$ décroît. Mais la constante $C$ reste indéterminée, ce qui peut être une source de complications. Voici l'une d'elles.

Supposons que la fonction $b$ de l'équation (2.40) oscille rapidement:

$$b(t) = f(t) \exp\left\{i\mu \int_0^t k(t)\, dt\right\}.$$

Formellement, la théorie exposée dans ce numéro nous fournit une recette pour construire une solution particulière, mais cette solution ne peut pas être utilisée, car les dérivées $d\tilde{y}^{(i)}/dt$ des seconds membres de (2.42) seront élevées, de l'ordre de $\mu$, si seulement $\mu$ est

*) Voir dans [7] la démonstration du théorème établissant les estimations (2.43).

assez grand. Il est évident que l'effet de $\mu$ disparaîtra lorsque $\lambda \to \infty$ mais dans la pratique on a toujours affaire à des $\lambda$ et $\mu$ finis.

Pour montrer à quelle sorte de particularités peut être confronté l'analyste, considérons une équation du second ordre avec un second membre oscillatoire, et, pour fixer les idées, supposons que $\mu = \lambda$:

$$\ddot{y} + 2a\dot{y} + \lambda^2\omega^2 y = \lambda^2 f(t) \exp\left\{i\lambda \int_0^t k(t)\,dt\right\}. \qquad (2.46)$$

Cherchons une solution particulière de l'équation (2.46) sous la forme

$$y = z(t, \lambda) \exp\left\{i\lambda \int_0^t k(t)\,dt\right\}.$$

On obtient l'équation suivante pour la fonction $z(t, \lambda)$:

$$\ddot{z} + 2i\lambda k(t)\dot{z} + i\lambda\dot{k}(t) z - \lambda^2 k^2 z + \\ + 2a(\dot{z} + i\lambda k(t) z) + \lambda^2\omega^2 z = \lambda^2 f(t). \qquad (2.47)$$

Supposons d'abord que $z(t, \lambda)$ est une fonction à variation lente. Autrement dit supposons qu'elle admet le développement

$$z(t, \lambda) = z_0(t) + \lambda^{-1} z_1(t) + \ldots$$

En portant ce développement dans l'équation (2.47), on obtient les équations suivantes pour les fonctions $z_i(t)$:

$$(\omega^2(t) - k^2(t))\, z_0 = f(t),$$
$$(\omega^2(t) - k^2(t))\, z_1 = -iz_0(\dot{k} - 2ak) - 2aik\dot{z}_0, \qquad (2.48)$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$

Les formules (2.48) permettent de calculer successivement tous les termes du développement de la fonction $z(t, \lambda)$ si seulement $k(t) \neq \omega(t)$ sur l'intervalle considéré. S'il existe au moins un point $t = t^* \in [0, T]$ en lequel $k(t^*) = \omega(t^*)$, alors cette procédure n'a plus de sens. Donc, dans ce cas, la fonction $z(t, \lambda)$ ne peut être traitée comme une fonction à variation lente. Mais cela ne signifie pas pour autant que l'on ne peut pas construire l'expression asymptotique de la solution particulière. Cette expression sera tout simplement différente, c'est-à-dire que la nature de la dépendance de la solution par rapport au paramètre sera différente. Pour construire cette nouvelle expression asymptotique, on se servira d'un procédé identique à celui qui nous a permis au chapitre précédent d'achever l'étude de la résonance principale. A cet effet, on admettra que le désaccord, c'est-à-dire la différence $\omega - k$, est une petite quantité

liée au paramètre $\lambda$ par la relation

$$\omega^2(t) - k^2(t) = \lambda^{-1}\chi(t). \tag{2.49}$$

Cherchons maintenant la fonction $z(t, \lambda)$ sous la forme

$$z(t, \lambda) = \lambda z_0 + z_1 + \lambda^{-1} z_2 + \ldots \tag{2.50}$$

En portant (2.49) et (2.50) dans l'équation (2.47) et en identifiant les coefficients des mêmes puissances de $\lambda$, on est conduit aux équations

$$\begin{aligned} 2ik\dot{z}_0 + (i\dot{k} + \chi + 2aki)\, z_0 &= f, \\ 2ik\dot{z}_1 + (i\dot{k} + \chi + 2aki)\, z_1 &= -\ddot{z}_0 - 2a\dot{z}_0, \end{aligned} \tag{2.51}$$

. . . . . . . . . . . . . . . . .

Chacune de ces équations est une équation différentielle ordinaire linéaire du premier ordre dont la solution peut être acquise par des quadratures. Ainsi, une solution particulière de la première équation du système (2.51) est

$$z_0 = \int_0^t \psi(\xi) \exp\left\{-\int_\xi^t \varphi(\zeta)\, d\zeta\right\} d\xi, \tag{2.52}$$

où

$$\psi(\xi) = \frac{f(\xi)}{2ik(\xi)}, \quad \varphi(\xi) = \frac{i\dot{k}(\xi) + \chi(\xi) + 2a(\xi)\, k(\xi)\, i}{2ik(\xi)}.$$

Si l'on se borne au premier terme du développement, on peut mettre une solution particulière de l'équation (2.46) sous la forme

$$y = \lambda z_0(t) \exp\left\{i\lambda \int_0^t k(t)\, dt\right\}. \tag{2.53}$$

REMARQUE. Si les fonctions de l'équation (2.46) sont des constantes, la formule (2.52) décrit alors les résonances dans des systèmes oscillatoires à paramètres constants. Pour s'en assurer on pose $\dot{k} = 0$ et on admet pour simplifier que $\chi = 0$, $a = 0$. Dans ces conditions $\varphi = 0$ et en tenant compte de ce que l'amplitude de la force perturbatrice $f$ est constante, on obtient la formule notoire des oscillations résonnantes:

$$y = \lambda \frac{f}{2i\omega} t \exp\{i\lambda k t\}.$$

En se servant de l'expression asymptotique de la solution particulière de l'équation (2.46), on intègre sans peine l'équation

$$\ddot{y} + 2a\dot{y} + \lambda^2\omega^2 y = \lambda^2 f(t) \cos\left\{\lambda \int_0^t k(t)\, dt\right\}.$$

20—01523

Mettons $\cos\left\{\lambda \int_0^t k(t)\,dt\right\}$ sous la forme

$$\cos\left\{\lambda \int_0^t k(t)\,dt\right\} = \frac{1}{2}\left[\exp\left\{i\lambda \int_0^t k(t)\,dt\right\} + \exp\left\{-i\lambda \int_0^t k(t)\,dt\right\}\right],$$

utilisons la linéarité du problème et posons $y = y_1 + y_2$, où $y_1$ et $y_2$ sont des solutions particulières des équations

$$\ddot{y} + 2a\dot{y} + \lambda^2\omega^2 y = \lambda^2 f^*(t) \exp\left\{i\lambda \int_0^t k(t)\,dt\right\},$$

$$\ddot{y} + 2a\dot{y} + \lambda^2\omega^2 y = \lambda^2 f^*(t) \exp\left\{-i\lambda \int_0^t k(t)\,dt\right\},$$

$$f^*(t) = f(t)/2.$$

Nous avons ainsi exhibé une procédure de construction des représentations asymptotiques des solutions particulières des équations différentielles avec seconds membres dans le cas où les forces extérieures sont des fonctions oscillatoires du temps et le système est à un degré de liberté. Signalons que tous les raisonnements effectués se généralisent à des systèmes linéaires arbitraires de la forme

$$\dot{y} = \lambda A(t, \lambda)\, y + f(t) \exp\left\{i\lambda \int_0^t k(t)\,dt\right\}, \tag{2.54}$$

où $y$ et $f$ sont des vecteurs de dimension $n$, $A$, la matrice

$$A(t, \lambda) = A_0(t) + \lambda^{-1}A_1(t) + \lambda^{-2}A_2(t) + \ldots$$

Si $k \neq \mu$ sur $[0, T]$, $\mu$ étant une racine de l'équation

$$|A_0 - \mu E| = 0,$$

alors la détermination d'une solution particulière de l'équation (2.54) se ramène à la résolution successive d'un système d'équations algébriques et la solution est de la forme

$$y = y_0(t) + \lambda^{-1}y_1(t) + \ldots \tag{2.55}$$

Si $k = \mu$ pour certains $t \in [0, T]$, alors on a affaire au cas résonnant et la détermination de la représentation asymptotique de la solution particulière réclame la résolution numérique d'un système d'équations différentielles: ce sont des équations linéaires de la forme (2.51) qui, à la différence de l'équation primitive, ne contiennent pas de termes à variation rapide. La représentation asymptoti-

que des solutions sera alors différente de (2.55) et sera de la forme

$$y = \lambda y_0 + y_1 + \lambda^{-1} y_2 + \ldots \tag{2.56}$$

e) *Fonctions frontières à variables lentes.* Revenons maintenant au problème déjà envisagé au n° b) de ce paragraphe. Nous allons utiliser la technique exposée au n° d) pour indiquer encore une méthode de construction des fonctions frontières pour les variables lentes.

Considérons les équations du système (2.1) et supposons connue la solution du système (2.2). Désignons cette solution par ($x^0$, $y^0$). D'après le schéma général, pour construire les fonctions frontières il faut linéariser le système (2.1) de la manière suivante:

$$\begin{aligned} \dot{x}_f &= \frac{\partial X}{\partial x} x_f + \frac{\partial X}{\partial y} y_f = B_1(t) x_f + B_2(t) y_f, \\ \dot{y}_f &= \lambda \frac{\partial Y}{\partial y} y_f + b_1(t) = \lambda A(t) y_f + b_1(t), \end{aligned} \tag{2.57}$$

où $b_1 = -dy^0/dt$. Considérons de nouveau le cas où $y$ est un scalaire; dans ces conditions $A(t) = -a(t)$ est une fonction scalaire et

$$y_f = \frac{b_1(t)}{\lambda a(t)} + C \exp\left\{-\lambda \int_0^t a(t)\, dt\right\} + O\left(\frac{1}{\lambda^2}\right).$$

En portant cette expression dans la première équation (2.57) et en négligeant les termes $O(1/\lambda^2)$, on obtient

$$\dot{x}_f = B_1(t) x_f + B_2(t) \left(\frac{1}{\lambda} \frac{b_1(t)}{a(t)} + C \exp\left\{-\lambda \int_0^t a\, dt\right\}\right), \tag{2.58}$$

où $C$ est une constante arbitraire qui peut être déterminée à partir des conditions initiales. La solution générale de l'équation (2.58) peut être mise sous la forme

$$x_f = x_f^*(t, C^*) + x_{f1} + x_{f2},$$

où $x_f^*(t, C^*)$ est la solution générale de l'équation homogène

$$\dot{x}_f^* = B_1(t) x_f^*.$$

Cette équation est d'intégration numérique aisée, car elle ne contient pas de variables rapides; $x_{f1}$ est une solution particulière de l'équation

$$\dot{x}_{f1} = B_1(t) x_{f1} + B_2(t) \frac{b_1}{\lambda a},$$

$x_{f2}$ une solution particulière de l'équation

$$\dot{x}_{f2} = B_1(t) x_{f2} + C B_2(t) \exp\left\{-\lambda \int_0^t a\, dt\right\}. \tag{2.59}$$

L'équation (2.59) peut être intégrée par des méthodes asymptotiques. Posons

$$x_{t2} = \exp\left\{-\lambda \int_0^t a\, dt\right\} z;$$

alors $z$ sera solution de l'équation

$$\dot{z} = \lambda a z + B_1 z + C B_2.$$

La représentation asymptotique d'une solution particulière sera

$$z = -\frac{CB_2}{\lambda a} - \frac{CB_2}{\lambda^2}\left[\frac{\dot{a}}{a^2} + \frac{1}{z} B_1 \frac{1}{a}\right] + \ldots,$$

et par suite

$$x_{t2} = -CB_2 \exp\left\{-\lambda \int_0^t a\, dt\right\}\left[\frac{1}{\lambda a} - \frac{1}{\lambda^2}\left(\frac{\dot{a}}{a^2} + \frac{1}{a} B_1 \frac{1}{a}\right)\right]. \qquad (2.60)$$

Cette méthode d'intégration approchée peut être étendue sans peine au cas général du système (2.1) où $y$ est un vecteur.

f) *Cas de racines multiples.* Dans les numéros précédents, nous avons vu que le changement du caractère de certains termes de l'équation se repercutait sur les propriétés de la solution approchée. Donc, en construisant les solutions approchées, il faut tenir compte de la structure des équations. En discutant les méthodes de construction de la solution générale des systèmes d'équations différentielles homogènes, on a admis que les racines de l'équation caractéristique étaient distinctes sur l'intervalle $[0, T]$ tout entier, c'est-à-dire qu'il n'existe aucun point $t = t^* \in [0, T]$ en lequel

$$\mu_i(t^*) = \mu_j(t^*), \qquad i \neq j. \qquad (2.61)$$

La condition $\mu_i(t) \neq \mu_j(t)$, $\forall\, t \in [0, T]$ peut être contraignante. Dans les problèmes d'ingénieur il arrive souvent que l'équation caractéristique possède des racines multiples. Rappelons par exemple que les fréquences du pendule sphérique étaient égales. Le problème des représentations asymptotiques dans le cas général est vaste et épineux. Nous ne l'étudierons pas en détail, nous contentant de quelques exemples illustrant les particularités que peut rencontrer l'analyste dans le cas de racines multiples.

Considérons le système homogène du second ordre

$$\begin{aligned} \dot{y}_1 + \lambda(a_{11} y_1 + a_{12} y_2) + b_{11} y_1 + b_{12} y_2 &= 0, \\ \dot{y}_2 + \lambda(a_{21} y_1 + a_{22} y_2) + b_{21} y_1 + b_{22} y_2 &= 0. \end{aligned} \qquad (2.62)$$

La structure des solutions approchées du système (2.62) est conditionnée par les singularités des racines de l'équation

$$\begin{vmatrix} a_{11}+\mu & a_{12} \\ a_{21} & a_{22}+\mu \end{vmatrix} = 0. \qquad (2.63)$$

Jusqu'ici nous avons traité le seul cas où les racines $\mu_1(t)$ et $\mu_2(t)$ étaient distinctes sur l'intervalle $[0, T]$ tout entier. Dans ce cas, le système (2.62) se ramène à la forme canonique suivante par une transformation linéaire:

$$\begin{aligned} \dot{z}_1 + \lambda\mu_1 z_1 + c_{11}z_1 + c_{12}z_2 &= 0, \\ \dot{z}_2 + \lambda\mu_2 z_2 + c_{21}z_1 + b_{22}z_2 &= 0. \end{aligned} \qquad (2.64)$$

Si désormais $\mu_1 = \mu_2$, on aura alors deux formes canoniques différentes correspondant à des structures différentes des diviseurs élémentaires.

α) Si les diviseurs élémentaires sont simples, alors le système peut être ramené à la forme (2.64), où $\mu_1 = \mu_2 = \mu$:

$$\begin{aligned} \dot{z}_1 + \lambda\mu z_1 + c_{11}z_1 + c_{12}z_2 &= 0, \\ \dot{z}_2 + \lambda\mu z_2 + c_{21}z_1 + c_{22}z_2 &= 0. \end{aligned} \qquad (2.65)$$

β) Si les diviseurs élémentaires ne sont pas simples, on ne peut plus réduire le système (2.62) à la forme (2.64). La forme canonique de ce système sera alors

$$\begin{aligned} z_1 + \lambda\mu z_1 + \lambda z_2 + c_{11}z_1 + c_{12}z_2 &= 0, \\ z_2 + \lambda\mu z_2 + c_{21}z_1 + c_{22}z_2 &= 0. \end{aligned} \qquad (2.66)$$

Il s'avère que les solutions approchées correspondant à ces deux cas sont fondamentalement différentes. Voyons tout d'abord le système (2.65). Cherchons sa solution sous la forme

$$z_1 = \exp\left\{-\lambda \int_0^t \mu\, dt\right\} x_1, \quad z_2 = \exp\left\{-\lambda \int_0^t \mu\, dt\right\} x_2. \qquad (2.67)$$

Ce changement élimine le paramètre $\lambda$ du système (2.65):

$$\begin{aligned} \dot{x}_1 + c_{11}x_1 + c_{12}x_2 &= 0, \\ \dot{x}_2 + c_{21}x_1 + c_{22}x_2 &= 0. \end{aligned} \qquad (2.68)$$

Ceci achève la construction de la représentation asymptotique du système d'équations (2.65): le système (2.68) ne contient plus le paramètre $\lambda$ et son intégration ne pose aucun problème, puisque les seconds membres ne contiennent pas de termes à variation rapide.

Désignons par $x_{1i}$ et $x_{2i}$, $i = 1, 2$, des solutions fondamentales du système (2.68). Ces fonctions peuvent être acquises par la résolution numérique de deux problèmes de Cauchy avec les conditions initiales suivantes:

$$\text{I. } x_{11}(0) = 0, \quad \text{II. } x_{12}(0) = 1,$$
$$x_{21}(0) = 1, \quad x_{22}(0) = 0.$$

La représentation asymptotique de l'intégrale générale du système (2.65), qui en l'occurrence est confondue avec la solution exacte, sera de la forme

$$\begin{aligned} z_1 &= \exp\left\{-\lambda \int_0^t \mu\, dt\right\}(c_1 x_{11} + c_2 x_{12}), \\ z_2 &= \exp\left\{-\lambda \int_0^t \mu\, dt\right\}(c_1 x_{21} + c_2 x_{22}). \end{aligned} \tag{2.69}$$

Considérons maintenant les équations (2.66). La méthode exposée ne convient visiblement pas, car le système obtenu par la transformation (2.67) contiendra les premières puissances du paramètre $\lambda$; les seconds membres de ce système seront à variation rapide et son intégration numérique posera des problèmes. On se servira néanmoins de la transformation (2.67) au premier pas. Ceci nous amène au système

$$\begin{aligned} \dot{x}_1 + \lambda x_2 + c_{11}x_1 + c_{12}x_2 &= 0, \\ \dot{x}_2 + c_{21}x_1 + c_{22}x_2 &= 0. \end{aligned} \tag{2.70}$$

Faisons encore un changement de variables: $x_1 = \lambda^{1/2}u$, $x_2 = v$. Ceci ramène le système (2.70) à la forme

$$\begin{aligned} \dot{u} + \lambda^{1/2}v + c_{11}u + \lambda^{-1/2}c_{12}v &= 0, \\ \dot{v} + \lambda^{1/2}c_{21}u + c_{22}v &= 0, \end{aligned} \tag{2.71}$$

Le système (2.71) fait partie des systèmes étudiés précédemment, car il renferme un nouveau grand paramètre $\lambda^* = \lambda^{1/2}$. On cherchera donc sa solution à l'aide du schéma classique:

$$u = \exp\left\{\lambda^{1/2} \int_0^t \omega\, dt\right\}\left(u_0 + \lambda^{-1/2} u_1 + \ldots\right),$$

$$v = \exp\left\{\lambda^{1/2} \int_0^t \omega\, dt\right\}\left(v_0 + \lambda^{-1/2} v_1 + \ldots\right),$$

où $\omega$ est une racine de l'équation caractéristique

$$\begin{vmatrix} \omega & 1 \\ c_{21} & \omega \end{vmatrix} = \omega^2 - c_{21} = 0. \tag{2.72}$$

Si $c_{21} \equiv 0$, la deuxième équation du système (2.71) s'intègre par des quadratures, et la solution générale de la première équation peut être exprimée par des quadratures, dont les intégrants seront uniquement des fonctions à variation lente. Si $c_{21} \not\equiv 0$, alors l'équation (2.72) admet deux racines distinctes $\omega_1$ et $\omega_2$. Les coefficients des développements de $u_i$ et $v_i$ se calculent explicitement par la méthode déjà décrite. En rassemblant les résultats, on peut mettre les deux solutions linéairement indépendantes du système (2.66) sous la forme suivante:

$$\begin{aligned} z_1^{(i)} &= \exp\left\{\int_0^t [-\lambda\mu + \lambda^{1/2}\omega_i]\,dt\right\}\left\{\lambda^{1/2}u_0^{(i)} + u_1^{(i)} + \ldots\right\}, \\ z_2^{(i)} &= \exp\left\{\int_0^t [-\lambda\mu + \lambda^{1/2}\omega_i]\,dt\right\}\left\{v_1^{(i)} + \lambda^{-1/2}v_1^{(i)} + \ldots\right\}. \end{aligned} \tag{2.73}$$

Les représentations asymptotiques des solutions approchées contiendront donc dans ce cas le paramètre $\lambda$ à des puissances fractionnaires. Ce fait (c'est-à-dire le fait que le développement contient des puissances fractionnaires lorsque les diviseurs élémentaires ne sont pas simples) a été de toute évidence signalé pour la première fois par Ya. Tamarkine (cf. [65]).

Les raisonnements qui ont servi à étudier le système du second ordre peuvent être généralisés à l'analyse d'un système arbitraire

$$y = A(t, \lambda)\, y, \tag{2.74}$$

où

$$A(t, \lambda) = \lambda^k (A^{k)}(t) + \lambda^{-1}A^{(k-1)}(t) + \ldots \\ \ldots + \lambda^{-s}A^{(k-s)}(t) + \lambda^{-s-1}H(t, \lambda)),$$

et la matrice $H(t, \;)$ est bornée pour $\lambda \to \infty$, $t \in [0, T]$. On appellera *rang* du système (2.74) la plus grande puissance $k$ du paramètre $\lambda$ et on admettra qu'une fonction $\mu(t)$ est une racine multiple de l'équation caractéristique $|A^{(k)} - \mu E| = 0$, mais que les diviseurs élémentaires de la matrice $A^{(k)}$ sont simples. La transformation

$$y(t, \lambda) = \exp\left\{\lambda^k \int_0^t \mu\, dt\right\} z(t, \lambda) \tag{2.75}$$

nous conduit au système

$$\dot{z}(t, \lambda) = B(t, \lambda) z(t, \lambda), \tag{2.76}$$

dont le rang sera inférieur d'une unité. Si le rang de la matrice $A^{(k)}$ était égal à l'unité, celui du système (2.76) serait nul, c'est-à-dire que le second membre de ce système ne contiendrait pas $\lambda$ à des puissances positives. Si, par exemple, la racine $\mu$ est double et les diviseurs élémentaires, simples, alors le système (2.76) doit être numériquement intégré. Ceci nous donnera deux solutions linéairement indépendantes $z_1$ et $z_2$ et de plus à la racine $\mu$ seront associées deux solutions particulières du système (2.75):

$$y_1 = \exp\left\{\lambda \int_0^t \mu \, dt\right\} z_1, \quad y_2 = \exp\left\{\lambda \int_0^t \mu \, dt\right\} z_2. \tag{2.77}$$

Si les diviseurs élémentaires ne sont pas simples, la situation peut être différente. Les puissances supérieures de $\lambda$ ne sont pas toutes éliminées par le changement de variables (2.75). Dans ce cas l'on est confronté à l'alternative suivante: ou bien les coefficients de l'équation (2.76) sont tels que les coefficients des termes de rang supérieur sont nuls par le changement (2.77), ou bien ils ne le sont pas. Dans le premier cas le rang du système diminue d'une unité et l'on peut poursuivre le processus de mise en évidence des facteurs exponentiels. Dans le deuxième cas on peut aussi construire les solutions asymptotiques, mais elles seront représentées par des séries suivant les puissances fractionnaires du paramètre $\lambda$.

g) *Remarques finales.* Dans ce paragraphe nous n'avons exposé qu'une seule méthode de construction des fonctions frontières, basée sur la linéarisation de l'équation par rapport à la variable rapide. Ceci étant, on a admis que l'écart entre la solution perturbée et la solution génératrice était petit, de même d'ailleurs que le desaccord des conditions initiales

$$y_0 - y^0(x_0, 0).$$

La dernière hypothèse peut limiter le domaine d'application de la procédure de calcul développée. Il existe aujourd'hui d'autres procédures plus générales de construction des fonctions frontières. Mais dans les calculs d'ingénieur on ne fait intervenir pour l'instant que les méthodes qui sont basées sur l'analyse des équations linéaires. Les algorithmes non linéaires sont encore mal étudiés sur le plan numérique.

Citons enfin un argument qui plaide particulièrement en faveur des théories linéaires des perturbations: il s'agit de la possibilité d'étudier les propriétés statistiques du système. Considérons encore

le système primitif

$$\dot{x} = X(x, y, t), \qquad \varepsilon\dot{y} = Y(x, y, t) \tag{2.78}$$

et sa solution génératrice $x^0(t)$, $y^0(t)$ qui est solution du système

$$\dot{x}^0 = X(x^0, y^0(x^0, t), t),$$

où $y^0$ est la racine de l'équation $Y(x^0, y^0, t) = 0$.

Nous pouvons souvent traiter la solution génératrice comme une solution de « base » définie par les conditions initiales fixes

$$x(0) = x_0.$$

Quant à la quantité $y_f = y - y^0(x^0, t)$, elle joue le rôle d'un « bruit de haute fréquence » engendré par le désaccord des conditions initiales

$$\delta = y_0 - y^0(x_0, 0).$$

L'utilisation des équations linéaires pour décrire les fonctions de couche limite $x_f$ et $y_f$ permet de traiter ces fonctions comme des opérateurs agissant sur le vecteur $\delta$. Pour tout $t$ on aura les expressions

$$y_f(t) = L_1\delta, \qquad x_f(t) = L_2\delta, \tag{2.79}$$

où $L_1$ et $L_2$ sont des matrices dont les éléments sont exprimés soit explicitement, soit par des quadratures élémentaires.

La simplicité de la forme des relations fonctionnelles $y_f(\delta)$ et $x_f(\delta)$ permet d'étudier les propriétés statistiques des trajectoires si l'on connaît celles des écarts initiaux $\delta$. Par exemple, le calcul de la matrice des moments d'ordre deux du vecteur $x_*(t)$ n'implique pas la résolution d'équations différentielles: on peut la déterminer explicitement.

L'appareil développé ici peut être appliqué à la résolution de problèmes plus compliqués où le système est soumis non seulement à des écarts initiaux aléatoires, mais aussi à l'action constante de perturbations aléatoires.

## § 3. Exemples de systèmes tikhonoviens et quasi tikhonoviens

Au § 1 nous avons étudié des systèmes de la forme

$$\dot{x} = X(x, y, t), \qquad \varepsilon\dot{y} = Y(x, y, t) \tag{3.1}$$

dans l'hypothèse que la solution de l'équation associée

$$dy/d\tau = Y(x, y, t), \tag{3.2}$$

où $x$ et $t$ sont des paramètres, est asymptotiquement stable. Ces systèmes ont été appelés systèmes tikhonoviens. Si, de plus, les

conditions initiales appartiennent au domaine d'attraction d'une racine $y^0$ de l'équation

$$Y(x, y^0, t) = 0, \tag{3.3}$$

on peut alors élaborer des méthodes d'approximation effectives pour l'analyse du système (3.1). Ces conditions peuvent certes être très astreignantes, notamment les conditions de stabilité asymptotique de la solution de l'équation (3.2), mais sans elles il est impossible d'établir des estimations asymptotiques, tout au moins pour des intervalles de temps assez grands.

Mais dans les problèmes pratiques, les intervalles de temps sont toujours bien définis, ce qui laisse espérer un relâchement de ces conditions de stabilité asymptotique. On a l'impression, par exemple, que pour réaliser les algorithmes décrits, il suffit d'exiger que les trajectoires du système associé donnent lieu à des écarts $y_t$ relativement petits sur l'intervalle de temps envisagé, si seulement les écarts initiaux $\delta$ sont assez petits.

Malheureusement, des considérations aussi vagues ne peuvent certainement pas servir d'assises à une quelconque théorie mathématique et à l'établissement d'estimations rigoureuses. Néanmoins la possibilité d'utiliser un appareil développé pour la résolution de problèmes pratiques moyennant des conditions plus faibles est extrêmement importante. De tels systèmes seront dits *quasi tikhonoviens*. Pour justifier les procédures développées pour ces systèmes, on ne peut se référer à aucun résultat mathématique : intuition, expérience et logique sont les seuls recours. Mais l'absence de résultats rigoureux n'estompe en aucun cas la valeur pratique des méthodes développées.

Dans ce paragraphe, on citera deux exemples considérés actuellement comme classiques. Ils sont consacrés à la mécanique du vol d'un obus et d'une fusée à empennage. L'étude de ces systèmes est très instructive en dépit de leur caractère assez spécial. Cette analyse souligne non seulement la différence entre les systèmes tikhonoviens et quasi tikhonoviens mais aussi l'unité de leur étude.

Quoique la théorie des systèmes tikhonoviens n'ait pris corps que dans les années cinquante, voire même dans les années soixante, les ingénieurs ont été confrontés dès le siècle dernier à des exemples de systèmes de la forme (3.1) pour lesquels ils ont élaboré des méthodes qui n'ont rien perdu de leur actualité.

Dans ce paragraphe nous abordons en particulier l'un des premiers problèmes de ce type : le problème général de balistique d'un obus tournant. La première tentative d'analyse du mouvement d'un projectile fut probablement entreprise dès les années quarante ou cinquante du siècle dernier par le général d'artillerie russe N. Maïevski qui étudia le mouvement relatif d'un projectile dans l'hypothèse que la trajectoire de son centre de masse est une droite. Mais c'est de Sparre [64] qui le premier se pencha sur ce problème dans sa

position complète. Dans ce paragraphe nous reproduisons les grandes lignes du schéma de son analyse en omettant les détails présentant un intérêt spécial. Pour simplifier les calculs on admettra que le projectile est complètement symétrique et qu'il est soumis à la seule action du moment extérieur des forces aérodynamiques (pour plus de détails voir [60, 73]).

Nous étudierons simultanément la mécanique du vol d'un projectile tournant et d'une fusée à empennage. Le mouvement du projectile se déroule dans l'espace, celui de la fusée peut, sous certaines conditions, être traité dans un plan (plan de tir).

a) *Mise en équation du mouvement.* Le mouvement du projectile est décrit par deux équations vectorielles : l'équation de la quantité

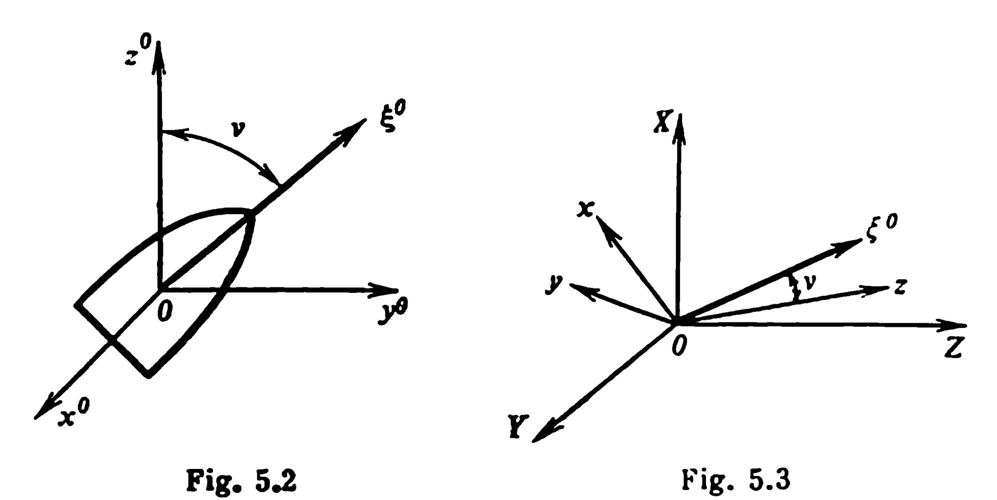



Fig. 5.2 Fig. 5.3

de mouvement et l'équation des moments. La première décrit le mouvement du centre de masse, la deuxième, le mouvement autour du centre de masse:

$$\frac{d^2r}{dt^2} = g + \frac{R}{m}, \tag{3.4}$$

$$\frac{d^2K}{dt^2} = L, \tag{3.5}$$

où $r$ est le rayon vecteur du centre de masse du projectile, $g$, le vecteur de gravitation terrestre, $R$, le vecteur forces extérieures, $K$, le vecteur moment cinétique, $L$, le vecteur moment des forces extérieures.

On admettra que les forces extérieures $R$ dépendent de la vitesse du centre de masse, de l'angle de nutation $\nu$ (fig. 5.2) et des dérivées de la vitesse de rotation angulaire du projectile. Le mouvement de la fusée est décrit par un système analogue d'équations, sauf qu'il faut adjoindre aux forces extérieures la poussée propulsive $P$ qui est proportionnelle à la perte de masse :

$$P = k\frac{dm}{dt}.$$

Considérant la perte de masse $dm/dt = q(t)$ (et partant la masse $m$) comme une fonction connue du temps, on peut étudier le mouvement de la fusée à l'aide du système (3.4), (3.5).

Les mouvements du projectile et de la fusée sont décrits en des termes qui n'ont pas la même signification. Ainsi, l'angle de nutation est appelé angle d'attaque en théorie des fusées, et ainsi de suite.

On distingue le problème fondamental et le problème général de balistique (dynamique). Le *problème fondamental de balistique extérieure* consiste à déterminer la trajectoire du centre de masse du projectile sous l'hypothèse que l'angle de nutation est nul. En d'autres termes, ce problème étudie le mouvement d'un projectile en l'assimilant à un point matériel. Ceci étant, la trajectoire est toujours plane si le projectile est symétrique.

On utilise d'autre part le terme de *problème général de balistique extérieure*. Ce problème étudie le mouvement autour du centre de masse, c'est-à-dire que le projectile ou la fusée est traité comme un solide.

La balistique extérieure part de l'hypothèse suivante: le mouvement autour du centre de masse peut être étudié avec une grande précision si l'on admet que le centre de masse se déplace le long d'une trajectoire qui est solution du problème fondamental de balistique extérieure. Cette hypothèse est essentielle, car elle permet d'étudier séparément les équations (3.4) et (3.5). Cette hypothèse a conditionné au XIX-XX-ièmes siècles le développement des méthodes mathématiques de balistique extérieure. Nous verrons plus bas qu'elle est la conséquence directe du fait que le système (3.4), (3.5) est proche d'un système tikhonovien dans des conditions réelles.

Le problème de balistique extérieure d'une fusée à empennage se formule exactement dans les mêmes termes. Mais à la différence du projectile, le système d'équations régissant le mouvement de la fusée est un système tikhonovien classique.

Ramenons maintenant le système (3.4), (3.5) à la forme scalaire. Introduisons à cet effet deux systèmes de coordonnées: un fixe et un mobile. Rattachons le système de coordonnées fixe $OXYZ$ au plan de tir, c'est-à-dire au plan qui contient la trajectoire du problème fondamental de balistique extérieure. Dirigeons l'axe $OX$ verticalement vers le haut (dans le sens contraire à celui de la force de pesanteur) et l'axe $OZ$, le long de l'intersection du plan de tir avec le plan horizontal. Disposons l'axe $OY$ dans le plan horizontal de telle sorte que le système $OXYZ$ soit direct (fig. 5.3).

Rattachons le système de coordonnées mobile $Oxyz$ au centre de masse du projectile (cf. fig. 5.2). Dirigeons l'axe $Oz$ le long du vecteur vitesse du centre de masse du projectile (resp. de la fusée). Les axes $Ox$ et $Oy$ sont situés dans un plan perpendiculaire à la direction de la vitesse (l'axe $Oz$). Lions leur sens au choix des axes du système fixe. Dirigeons l'axe $Ox$ le long de l'intersection du plan vertical $XOY$

et du plan normal à l'axe $Oz$; l'axe $Oy$, perpendiculairement aux axes $Ox$ et $Oz$ de telle sorte que le système $Oxyz$ soit direct. Donc l'axe $Oy$ est contenu dans le plan horizontal $YOZ$. Le système de coordonnées $Oxyz$ est mobile et tourne par rapport au système de coordonnées fixe avec une vitesse angulaire instantanée $\omega^*$. Si le centre de masse se déplaçait toujours dans le plan de tir $XOZ$ et par suite son vecteur vitesse était dans ce plan, la vitesse angulaire $\omega^*$ serait nulle. Donc, le système mobile introduit ne participe que dans le mouvement de la tangente à la trajectoire du projectile et pas dans son mouvement propre. Formons maintenant l'équation du moment cinétique par rapport au système mobile. Désignons à cet effet par $d\tilde{K}/dt$ la dérivée du moment cinétique $K$ par rapport au système mobile $Oxyz$. L'équation (3.5) devient alors

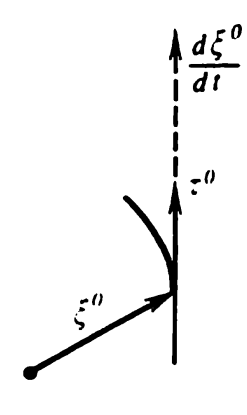


Fig. 5.4

$$\frac{d\tilde{K}}{dt}+\omega^*\times K=L. \tag{3.6}$$

Désignons par $\Omega$ le vecteur vitesse angulaire instantanée du projectile, par $\omega$ le vecteur vitesse de la rotation propre du projectile autour de son axe de symétrie de vecteur directeur unitaire $\xi^0$. Alors

$$\omega = |\omega|\,\xi^0, \tag{3.7}$$

où $|\omega|$ est le module du vecteur $\omega$. Posons

$$\Omega = \omega + \omega_1. \tag{3.8}$$

Etudions plus en détail le vecteur $\omega_1$. Le mouvement de rotation du projectile peut être décomposé en deux mouvements: une rotation propre autour de l'axe de symétrie de vecteur directeur $\xi^0$ et une rotation du vecteur $\xi^0$. Donc, $\omega_1$ est la vitesse angulaire instantanée du vecteur $\xi^0$. La quantité $d\xi^0/dt$ est la vitesse de l'extrémité du vecteur unitaire $\xi^0$: $d\xi^0/dt = \omega_1 \times \xi^0$. Or $\xi^0$ est un vecteur unitaire et $\omega_1$, sa vitesse angulaire instantanée, c'est-à-dire que $\omega_1$ est orthogonal à $\xi^0$. Donc $d\xi^0/dt = |\omega_1|\,\tau^0$, où $\tau^0$ est le vecteur unitaire tangent à l'hodographe du vecteur $\xi^0$ (fig. 5.4) et $|\omega_1|$ le module du vecteur $\omega_1$. Comme $\tau^0$ est unitaire, on a

$$|\omega_1| = \left|\frac{d\xi^0}{dt}\right|. \tag{3.9}$$

Le vecteur $\omega_1$ est perpendiculaire au plan passant par le vecteur $\xi^0$ et le vecteur vitesse linéaire de son extrémité. Donc, le vecteur $\omega_1$ est de même sens que vecteur

$$\xi^0\times\frac{d\xi^0}{dt}. \tag{3.10}$$

D'autre part les deux vecteurs de ce produit sont orthogonaux. Mais

$$\left|\xi^0 \times \frac{d\xi^0}{dt}\right| = |\xi^0| \left|\frac{d\xi^0}{dt}\right| \sin\varphi,$$

où $\varphi = \pi/2$ et $|\xi^0| = 1$. Donc,

$$\left|\frac{d\xi^0}{dt}\right| = \left|\xi^0 \times \frac{d\xi^0}{dt}\right|.$$

En comparant cette égalité avec (3.9), on voit que

$$|\omega_1| = \left|\xi^0 \times \frac{d\xi^0}{dt}\right|.$$

Le vecteur $\omega_1$ étant non seulement de même module mais aussi de même sens que (3.10), on a

$$\omega_1 = \xi^0 \times \frac{d\xi^0}{dt}. \tag{3.11}$$

Les expressions (3.7) et (3.8) aidant, on peut mettre $\Omega$ sous la forme

$$\Omega = |\omega| \xi^0 + \xi^0 \times \frac{d\xi^0}{dt}. \tag{3.12}$$

Nous avons écrit l'expression de $\omega_1$ dans le système fixe. Passons maintenant au système mobile; il faut à cet effet remplacer $d\xi^0/dt$ par son expression

$$\frac{d\xi^0}{dt} = \frac{d\tilde{\xi}^0}{dt} + \omega^* \times \xi^0,$$

où $\tilde{\xi}^0$ désigne le vecteur $\xi^0$ par rapport au système de coordonnées mobile. Comme $\xi^0 \times (\omega^* \times \xi^0) = \omega^* - \xi^0 (\omega^*, \xi^0)$, l'expression (3.11) devient

$$\omega_1 = \xi^0 \times \frac{d\tilde{\xi}^0}{dt} + \omega^* - \xi^0 (\xi^0, \omega^*). \tag{3.13}$$

Composons maintenant l'expression du moment cinétique

$$K = J\Omega,$$

où $J$ est le tenseur d'inertie. Le projectile est par hypothèse à symétrie axiale. Donc, ses caractéristiques sont déterminées par deux moments d'inertie: le moment équatoriel $A$ et le moment polaire $C$. Il est évident alors que

$$K = A\omega_1 + C\omega. \tag{3.14}$$

Utilisons maintenant les formules (3.14) et (3.13) et transformons l'équation (3.6):

$$A\left\{\xi^0 \times \frac{d^2\tilde{\xi}^0}{dt^2} - \frac{d\xi^0}{dt}(\xi^0, \omega^*) - \xi^0\left(\frac{d\tilde{\xi}^0}{dt}, \omega^*\right) - \right.$$

$$\left. - \xi^0\left(\xi^0, \frac{d\omega^*}{dt}\right) + \frac{d\tilde{\omega}^*}{dt}\right\} + C\left(\frac{d\,|\,\omega\,|}{dt}\xi^0 + |\,\omega\,|\frac{d\tilde{\xi}^0}{dt}\right) +$$

$$+ \omega^* \times \left\{C\,|\,\omega\,|\,\xi^0 + A\left[\xi^0 \times \frac{d\tilde{\xi}^0}{dt} - \xi^0(\xi^0, \omega^*)\right]\right\} = L. \quad (3.15)$$

En rassemblant les termes semblables, on obtient

$$C\left(\frac{d\,|\,\omega\,|}{dt}\xi^0 + |\,\omega\,|\frac{d\tilde{\xi}^0}{dt} + |\,\omega\,|(\omega^* \times \xi^0)\right) +$$

$$+ A\left\{\xi^0 \times \frac{d^2\tilde{\xi}^0}{dt^2} - 2\frac{d\tilde{\xi}^0}{dt}(\xi^0, \omega^*) - \xi^0\left(\xi^0, \frac{d\tilde{\omega}^0}{dt}\right) + \right.$$

$$\left. + \frac{d\omega^*}{dt} - (\xi^0, \omega^*)(\omega^* \times \xi^0)\right\} = L. \quad (3.16)$$

Le moment $L$ est le moment des forces aérodynamiques. On peut le représenter avec une grande précision par la somme de deux termes: $L = L_1 + L_2$, où $L_1$ dépend de l'angle de nutation et de ses dérivées et $L_2$, de la vitesse de rotation propre. En se bornant à l'approximation linéaire, on met généralement ces quantités sous la forme

$$L_1 = \varkappa(t)(z^0 \times \xi^0) + k_1(t)\frac{d}{dt}(z^0 \times \xi^0), \quad (3.17)$$

$$L_2 = -k_2(t)\,\omega = -k_2(t)\,|\,\omega\,|\,\xi^0. \quad (3.18)$$

La première composante de la somme du second membre de (3.17) s'appelle moment basculateur: si le projectile ne tournait pas, il basculerait sous l'action des forces aérodynamiques. Sur la figure 5.5, le point $O$ désigne le centre de masse, le point $A$, le point d'application des forces aérodynamiques (le foyer aérodynamique). La rotation propre stabilise le vol du projectile. Le moment aérodynamique est dirigé dans un autre sens pour la fusée: la présence du stabilisateur déplace le foyer aérodynamique au-delà du centre de masse $O$ (fig. 5.6), de sorte que le moment contrecarrera la croissance de l'angle de nutation (angle d'attaque) comme l'indique la flèche de la figure 5.6.

La deuxième composante de l'équation (3.17) s'appelle moment amortissant. Elle dépend de la vitesse de variation de l'angle de nutation et elle est toujours opposée à la vitesse de variation de l'angle de nutation. En d'autres termes, le coefficient $k_1$ est toujours strictement négatif. Pour simplifier les calculs on négligera le moment amortissant, c'est-à-dire on posera $k_1 = 0$. Le moment amortis-

sant joue exactement le même rôle que le frottement. Nous utiliserons cette circonstance dans la suite.

Donc, $k_1 = 0$. De l'équation (3.16) on déduit alors immédiatement une équation scalaire décrivant les variations de la vitesse angulaire $\omega$. En effet, en multipliant les deux membres de l'équation (3.16) scalairement par $\xi^0$, on obtient

$$C \frac{d\,|\,\omega\,|}{dt} = -k_2\,|\,\omega\,|,$$

d'où

$$|\,\omega\,| = \omega_0 \exp\left(-\frac{k_2}{C}t\right). \tag{3.19}$$

On peut donc étudier le mouvement de rotation du projectile autour de son axe indépendamment de l'autre mouvement. Le moment

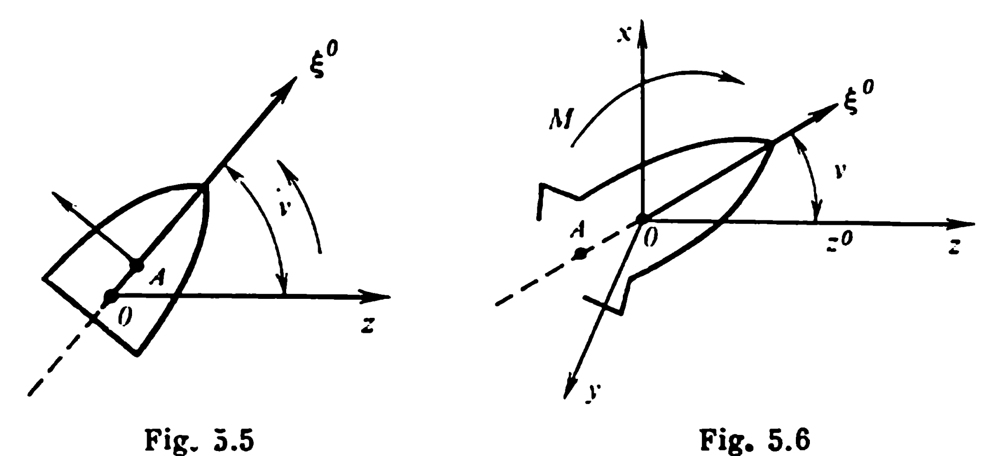



Fig. 5.5 Fig. 5.6

$k_2\,|\,\omega\,|$ est le moment de frottement des forces aérodynamiques. L'expression (3.19) nous dit que la vitesse de rotation angulaire tend progressivement vers 0. Si $k_2 = 0$, alors

$$\omega = \text{const}, \tag{3.20}$$

c'est-à-dire que dans ces conditions le système étudié admet (3.20) pour intégrale première. La fusée à empennage ne tourne pas et à tout instant on a

$$\omega = 0. \tag{3.20'}$$

Etant donné que nous avons admis que la fusée est à symétrie axiale, la poussée propulsive ne crée pas de moment et le mouvement de la fusée par rapport à son centre de masse est décrit par l'équation (3.16) dans laquelle il faut poser $\omega = 0$.

REMARQUE. Vu que nous avons négligé l'action du moment amortissant, il est naturel d'en faire autant pour la diminution de la vitesse de rotation angulaire due au frottement de l'air, puisque dans les deux cas il est question de forces de frottement. Remarquons seulement que les simplifications faites tout au long de l'exposé ne restreignent en rien les particularités du phénomène étudié, elles ne font que réduire le volume des calculs.

Supposons donc que $d\omega/dt = 0$. Le mouvement relatif du projectile admet trois degrés de liberté. Nous venons tout juste de parler de l'un d'eux. Pour décrire entièrement le mouvement relatif du projectile il nous faut donc déduire encore deux équations scalaires à partir de l'équation vectorielle (3.16). Ecrivons cette équation en projection sur les axes $Ox$ et $Oy$ du système de coordonnées mobile. Désignons par $x$, $y$ et $z$ les projetés du vecteur $\xi^0$ sur les axes du système mobile. Les projetés de $\tilde{d}\xi^0/dt$ sur les axes $Ox$ et $Oy$ seront respectivement $dx/dt$ et $dy/dt$. Les projetés du produit vectoriel $\xi^0 \times \frac{\tilde{d}^2\xi^0}{dt^2}$ sur ces axes seront

$$y\frac{d^2z}{dt^2} - z\frac{d^2y}{dt^2} \quad \text{et} \quad z\frac{d^2x}{dt^2} - x\frac{d^2z}{dt^2}.$$

Considérons le vecteur $\omega^*$; le système mobile peut effectuer une rotation seulement autour des axes $Ox$ et $Oy$, donc

$$\omega^* = \omega_x^* x^0 + \omega_y^* y^0,$$

où $x^0$ et $y^0$ sont les vecteurs unitaires respectifs des axes $Ox$ et $Oy$. La rotation autour de l'axe $Ox$ fait varier l'angle $\psi$ formé par l'axe $Oz$ (le vecteur vitesse) et le plan vertical. Convenons de mesurer cet angle à partir du plan vertical. Alors

$$\omega_x^* = -\dot{\psi}.$$

De façon analogue

$$\omega_y^* = \dot{\theta},$$

où $\theta$ est l'angle du vecteur vitesse avec l'horizontale. Ainsi

$$\omega^* = -\dot{\psi}x^0 + \dot{\theta}y^0.$$

Nous pouvons maintenant calculer sans peine les projetés des autres vecteurs et les produits figurant dans l'équation (3.16):

$$\omega^* \times \xi^0 = \dot{\theta}zx^0 + \dot{\psi}zy^0 - (\dot{\psi}y + \dot{\theta}x)\,z^0,$$

$$\frac{\tilde{d}\omega^*}{dt} = -\ddot{\psi}x^0 + \ddot{\theta}y^0, \quad (\xi^0, \omega^*) = -\dot{\psi}x + \dot{\theta}y,$$

$$\left(\xi^0, \frac{\tilde{d}\omega^*}{dt}\right) = -\ddot{\psi}x + \ddot{\theta}y, \quad (z^0 \times \xi^0) = -yx^0 + xy^0.$$

Multiplions maintenant l'équation (3.16) successivement par $x^0$ et $y^0$. En se servant des expressions auxiliaires obtenues, on est

conduit aux deux équations scalaires suivantes:

$$A(y\ddot{z} - \ddot{y}z) + C\omega\dot{x} - 2A\dot{x}(\dot{\theta}y - \dot{\psi}x) - Ax(\ddot{\theta}y - \ddot{\psi}x) -$$

$$- A\ddot{\psi} - C\omega\dot{\theta}z - A\dot{\theta}z(\dot{\theta}y - \dot{\psi}x) = -\varkappa(t)\, y, \qquad (3.21)$$

$$A(\ddot{x}z - \ddot{z}x) + C\omega\dot{y} - 2A\dot{y}(\dot{\theta}y - \dot{\psi}x) - Ay(\ddot{\theta}y - \ddot{\psi}x) +$$

$$+ A\ddot{\theta} + C\omega\dot{\psi}z - A\dot{\psi}z(\dot{\theta}y - \dot{\psi}x) = \varkappa(t)\, x.$$

Ces deux équations combinées à l'intégrale (3.20) et la condition de normalisation

$$x^2 + y^2 + z^2 = 1$$

décrivent entièrement le mouvement du projectile ou de la fusée autour du centre de masse.

Considérons maintenant le système (3.4) et mettons-le sous la forme suivante

$$\frac{dr}{dt} = v, \qquad (3.22)$$

$$\frac{dv}{dt} = \left|\frac{dv}{dt}\right| z^0 + |v| \frac{dz^0}{dt} = g - \frac{R}{m}, \qquad (3.23)$$

où $v = |v|\, z^0$ est la vitesse du mouvement du centre de masse. Dans le cas du projectile, $R$ est la résistance de l'air. Elle dépend de la vitesse $v$ du centre de masse et de l'angle de nutation $\nu$:

$$R = R(v, \nu) = R(v, x, y).$$

Si nous étudions le mouvement d'une fusée à empennage, alors l'équation (2.23) doit être mise sous la forme

$$\frac{dv}{dt} = g + \frac{R}{m} + \frac{P}{m}, \qquad (3.23)$$

où $P$ est la poussée des réacteurs; de même que la masse, la quantité $P$ est une fonction du temps supposée connue. Elle est orientée le long de l'axe de la fusée:

$$P = -|P(t)|\, \xi^0,$$

et l'équation (3.23) sera mise sous la forme

$$\frac{dv}{dt} = g + \frac{R}{m} = f(t)\, \xi^0, \qquad (3.23'')$$

où

$$f(t) = \frac{|P(t)|}{m(t)}.$$

Ecrivons l'équation (3.22) en projection sur les axes du système de coordonnées fixe.

Les angles $\theta$ et $\psi$ sont représentés sur la figure 5.7. En désignant donc par $X$, $Y$ et $Z$ les projetés du rayon vecteur du centre de masse par rapport au système de coordonnées fixe, on obtient les équations scalaires (ici et plus bas $v = |v|$)

$$\frac{dX}{dt} = v \sin \theta, \qquad \frac{dY}{dt} = v \sin \psi. \tag{3.24}$$

Comme $\left(\frac{dZ}{dt}\right)^2 + \left(\frac{dY}{dt}\right)^2 + \left(\frac{dX}{dt}\right)^2 = v^2$, on obtient en vertu de (3.24) l'équation suivante pour $Z$:

$$\frac{dZ}{dt} = v \sqrt{\cos^2 \theta - \sin^2 \psi}. \tag{3.25}$$

Les équations scalaires (3.24) et (3.25) sont équivalentes à l'équation vectorielle (3.22).

Mettons maintenant l'équation (3.23) sous la forme scalaire. A cet effet multiplions scalairement ses deux membres par le vecteur $z^0$

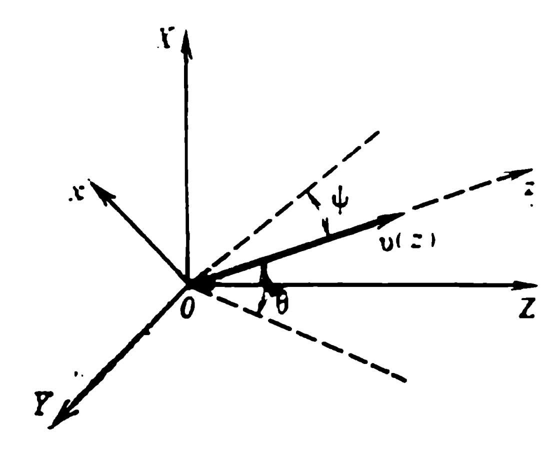



Fig. 5.7

(le vecteur directeur de la vitesse). Comme $dz^0/dt$ est orthogonal à $z^0$ et $g = -|g| X^0$ ($X^0$ est le vecteur unitaire de $OX$), il vient

$$\frac{dv}{dt} = -g \sin \theta + \frac{R(x, y, X)}{m} = -g \sin \theta + R_1. \tag{3.26}$$

Ecrivons maintenant l'équation (3.23) en projection sur les axes $Ox$ et $Oy$ (fig. 5.7):

$$\begin{aligned} \frac{d\theta}{dt} &= -\frac{g \cos \theta}{v} + \frac{Q_1(x, y, X)}{mv} = -\frac{g \cos \theta}{v} + R_2, \\ \frac{d\psi}{dt} &= \frac{Q_2(x, y, X)}{mv} = R_3. \end{aligned} \tag{3.27}$$

Les équations (3.26) et (3.27) sont équivalentes à l'équation vectorielle (3.23) et $R_1$, $R_2$, $R_3$ sont les projetées des accélérations engendrées par la force aérodynamique sur les axes du système de coordonnées

mobile. Ces projetées sont des fonctions de toutes les variables principales du problème:

$$R_i = R_i\,(v,\ m,\ x,\ y,\ X). \tag{3.28}$$

La fonction $R$ s'appelle résistance frontale, $Q_1$ et $Q_2$ sont les composantes de la portance. Ces composantes s'annulent pour $x = y = 0$ et peuvent être représentées avec une grande précision par les formules

$$\frac{Q_1}{m} + qx, \quad \frac{Q_2}{m} = qy. \tag{3.28'}$$

Si l'on étudie le mouvement d'une fusée, il faut inclure dans les équations les composantes de la poussée propulsive. En se servant de la formule (3.28'), on écrira les équations (3.26), (3.27) sous la forme suivante pour le cas de la fusée:

$$\frac{dv}{dt} = -g \sin\theta + R_1 - f(t)\,z, \tag{3.26'}$$

$$\left.\begin{aligned} \frac{d\theta}{dt} &= \frac{-g\cos\theta}{v} + R_2 - f(t)\,x, \\ \frac{d\psi}{dt} &= R_3 - f(t)\,y. \end{aligned}\right\} \tag{3.27'}$$

Ainsi le mouvement d'un projectile tournant ou d'une fusée est régi par le système d'équations (3.21), (3.24) à (3.27), (3.26') et (3.27'). C'est un système complexe d'équations différentielles non linéaires qui pose des difficultés pour l'analyse numérique, puisque le mouvement du projectile (ou de la fusée) est un mouvement oscillatoire à trois dimensions de haute fréquence.

b) *Introduction d'un paramètre et approximation d'ordre zéro.* Le système d'équations composé se rapporte sous certaines conditions aux systèmes tikhonoviens ou quasi tikhonoviens. Pour s'en assurer il faut introduire un paramètre dans ce système d'équations et le ramener à la forme standard (3.1). Cette réduction sera basée sur certaines particularités du mouvement du projectile ou de la fusée, engendrant une oscillation rapide d'une partie des coordonnées de phase. Les mouvements du projectile et de la fusée seront étudiés séparément. Commençons par le premier.

Le mouvement de rotation et d'oscillation à haute fréquence du projectile autour de son centre de masse est dû à l'action de deux facteurs: premièrement, il existe un grand moment basculateur des forces aérodynamiques et, deuxièmement, un effet gyroscopique compensant le moment basculateur. En théorie élémentaire du gyroscope il existe une relation de stabilité du mouvement, reliant le coefficient du moment basculateur $\varkappa$ au moment gyroscopique $C\omega^2$:

$$C\omega^2 > \varkappa.$$

Donc, pour que le mouvement soit stable, il est nécessaire que la vitesse angulaire $\omega$ de rotation propre du gyroscope soit assez grande:

$$\omega \geqslant \hat{\omega} = \sqrt{\frac{\varkappa}{C}}, \tag{3.29}$$

c'est-à-dire que $\omega^2$ et $\varkappa$ doivent être du même ordre. Posons

$$\omega = \lambda\mu, \quad \varkappa = \lambda^2 n^2 A, \tag{3.30}$$

où $\lambda$ est un grand paramètre. Introduisons par ailleurs les nouvelles variables

$$\dot{x} = \lambda\alpha, \quad \dot{y} = \lambda\beta, \quad \dot{z} = \lambda\gamma. \tag{3.31}$$

La signification du changement de variables (3.31) est évidente: nous étudions les mouvements d'un projectile dont l'angle de nutation varie avec une vitesse élevée.

Posons $\varepsilon = 1/\lambda$. Il vient alors

$$\varepsilon\dot{x} = \alpha, \qquad \varepsilon\dot{y} = \beta, \qquad \varepsilon\dot{z} = \gamma. \tag{3.32}$$

En se servant des notations (3.32), on peut mettre les équations (3.21) sous la forme suivante:

$$\begin{gathered}
\varepsilon\{A(y\dot{\gamma} - z\dot{\beta}) - 2A\alpha(\dot{\theta}y - \dot{\psi}x) - \varepsilon Ax(\ddot{\theta}y - \ddot{\psi}x) - \\
- \varepsilon A\ddot{\psi} - C\mu\dot{\theta}z - \varepsilon A\dot{\theta}z(\dot{\theta}y - \dot{\psi}x)\} = -An^2y - C\mu\alpha, \\
\varepsilon\{A(\dot{\alpha}z - \dot{\gamma}x) - 2A\beta(\dot{\theta}y - \dot{\psi}x) - \varepsilon Ay(\ddot{\theta}y - \ddot{\psi}x) + \\
+ \varepsilon A\ddot{\theta} + C\mu\dot{\psi}z - \varepsilon A\dot{\psi}z(\dot{\theta}y - \dot{\psi}x)\} = An^2x - C\mu\beta.
\end{gathered} \tag{3.33}$$

Pour ramener le système (3.33) à la forme standard il faut encore éliminer $z$ et $\gamma$ à l'aide de la condition de normalisation $x^2 - y^2 + z^2 = 1$ et résoudre le système obtenu par rapport aux dérivées, $\dot{\alpha}$ et $\dot{\beta}$. Ceci nous conduit à des équations de la forme (3.1). Mais nous renoncerons à cette procédure, car le système (3.32), (3.33) est assez simple en soi.

En posant $\varepsilon = 0$, on obtient

$$\alpha = 0, \quad \beta = 0, \quad An^2y + C\mu\alpha = 0, \quad An^2x - C\mu\beta = 0,$$

d'où il s'ensuit qu'en approximation d'ordre zéro $x = y = 0$ et $z = 1$, c'est-à-dire que l'axe du projectile est confondu avec la tangente à la trajectoire. Ceci revient à supposer que le mouvement relatif du projectile peut être négligé en approximation d'ordre zéro. Donc, pour $\varepsilon = 0$ ($\lambda = \infty$), on est conduit au problème fondamental de balistique extérieure qui consiste à calculer la trajectoire du cen-

tre de masse sous l'hypothèse que le projectile est assimilé à un point matériel.

Donc, si les valeurs initiales des variables $x$, $y$, $\dot{x}$, $\dot{y}$ appartiennent au domaine d'attraction de la racine $x = y = \dot{x} = \dot{y} = 0$ et si le mouvement du projectile est stable, alors le système d'équations régissant ce mouvement est tikhonovien et la solution d'approximation d'ordre zéro, c'est-à-dire la solution du problème fondamental de balistique extérieure, approche uniformément la trajectoire du centre de masse du projectile à $O(1/\lambda)$ près.

En faisant coïncider le plan de tir avec le plan $XOZ$, c'est-à-dire en supposant que $\psi(0) = 0$, on trouve à partir de la deuxième équation du système (3.27) que $\psi(t) = 0$ en approximation zéro. De là il s'ensuit que $Y = 0$ en approximation zéro, c'est-à-dire que la trajectoire génératrice est plane.

Les raisonnements sont identiques pour le cas de la fusée. On peut utiliser les équations (3.21) pour décrire le mouvement de la fusée par rapport à son centre de masse. Il suffit à cet effet de poser $\omega = 0$ et de tenir compte du fait que le moment des forces aérodynamiques n'est plus basculateur, mais de rappel. Il faut donc changer le signe de $\varkappa$ dans les équations (3.21). Ceci nous conduit au système d'équations suivant :

$$\begin{gathered}
A(y\ddot{z} - \ddot{y}z) - 2A\dot{x}(\dot{\theta}y - \dot{\psi}x) - Ax(\ddot{\theta}y - \ddot{\psi}x) - \\
- A\ddot{\psi} - A\dot{\theta}z(\dot{\theta}y - \dot{\psi}x) = \varkappa(t)y, \\
A(\ddot{x}z - \ddot{z}x) - 2A\dot{y}(\dot{\theta}y - \dot{\psi}x) - Ay(\ddot{\theta}y - \ddot{\psi}x) + \\
+ A\ddot{\theta} + C\dot{\psi}z - A\dot{\psi}z(\dot{\theta}y - \dot{\psi}x) = -\varkappa(t)x.
\end{gathered} \tag{3.34}$$

Le mouvement de la fusée à empennage par rapport à son centre de masse est un mouvement oscillatoire à haute fréquence d'un pendule sphérique soumis à l'action d'un moment aérodynamique de rappel élevé. En vertu de cela, introduisons de nouveau le paramètre

$$\varkappa = \lambda^2 n^2 A$$

et faisons le changement (3.31). On est amené en définitive aux équations

$$\begin{gathered}
\varepsilon\{A(y\dot{\gamma} - z\dot{\beta}) - 2A\alpha(\dot{\theta}y - \dot{\psi}x) - \varepsilon Ax(\ddot{\theta}y - \ddot{\psi}x) - \\
- \varepsilon A\ddot{\psi} - \varepsilon A\dot{\theta}z(\dot{\theta}y - \dot{\psi}x)\} = An^2y, \\
\varepsilon\{A(\dot{\alpha}z - \dot{\gamma}x) - 2A\beta(\dot{\theta}y - \dot{\psi}x) - \varepsilon A(\ddot{\theta}y - \ddot{\psi}x) + \\
+ \varepsilon A\ddot{\theta} - \varepsilon A\dot{\psi}z(\dot{\theta}y - \dot{\psi}x)\} = -An^2x.
\end{gathered} \tag{3.35}$$

En faisant $\varepsilon = 0$ dans les équations (3.32) et (3.35), on trouve que $\alpha = 0$, $\beta = 0$, $x = 0$ et $y = 0$ sur la solution génératrice et l'on aboutit aux mêmes conclusions que lors de l'analyse du mouvement du projectile: en approximation zéro, la trajectoire de la fusée est une courbe plane et l'axe de la fusée est confondu avec le support de la vitesse du centre de masse.

c) *Déduction des systèmes d'équation associés de la mécanique du vol de la fusée et du projectile tournant.* Les équations associées sont, suivant la terminologie de Tikhonov, les équations du mouvement (relatif) de rotation (3.33) (ou (3.35)) dans lesquelles les variables $X$, $Y$, $Z$, $v$, $\theta$ et $\psi$ sont traitées comme des paramètres pris égaux à leurs valeurs sur la trajectoire génératrice, c'est-à-dire sur la trajectoire définie par la solution du problème fondamental de balistique extérieure. C'est un système de deux équations du second ordre par rapport à $x$ et $y$, projetés du vecteur unitaire de l'axe du projectile sur les axes du système de coordonnées mobile. Les coefficients de ce système dépendent de la solution du système générateur et par suite sont des fonctions du temps connues. Pour le résoudre approximativement on supposera que les $x$ et $y$ sont de petites quantités et alors le système (3.33) peut être remplacé par un système linéaire. Admettre que $x$ et $y$ sont petites revient à admettre que l'angle de nutation l'est aussi: cette hypothèse est non seulement naturelle mais elle est toujours réalisée dans les constructions réelles. Si l'angle de nutation d'un projectile (ou l'angle d'attaque pour une fusée) est élevé, alors le mouvement ne peut pratiquement pas être stable.

A noter qu'admettre que des quantités sont petites, revient à introduire un nouveau paramètre indépendant du paramètre $\varepsilon$ envisagé au n° b). Donc, en désignant par $\delta$ une quantité petite, on convient que

$$x = O(\delta), \qquad y = O(\delta). \tag{3.36}$$

Il est alors évident (en vertu de la normalisation $x^2 + y^2 + z^2 = 1$) que $z = \sqrt{1 - x^2 - y^2} = 1 - O(\delta^2)$. On admettra aussi que les dérivées $\dot{x}$ et $\dot{y}$ sont des quantités petites du premier ordre:

$$\dot{x} = O(\delta), \ \dot{y} = O(\delta).$$

Alors

$$\dot{z} = \frac{\partial z}{\partial x}\dot{x} + \frac{\partial z}{\partial y}\dot{y} = O(\delta^2). \quad \ddot{z} = O(\delta^2).$$

On supposera encore que le système de coordonnées $Oxyz$ tourne lentement, c'est-à-dire que les dérivées de la vitesse angulaire sont également petites:

$$\dot{\theta} = O(\delta), \qquad \dot{\psi} = O(\delta).$$

En se servant de ces estimations et en négligeant dans les équations (3.21) les termes d'ordre $O(\delta^2)$ et plus, on obtient

$$A\ddot{x} + C\omega\dot{y} - \varkappa x = -A\ddot{\theta} - C\omega\dot{\psi},$$
$$A\ddot{y} - C\omega\dot{x} - \varkappa y = -C\omega\dot{\theta} - A\ddot{\psi}$$

ou, en introduisant le paramètre $\lambda$ conformément aux égalités (3.30), on met ces équations sous la forme

$$\begin{aligned} A\ddot{x} + C\lambda\mu\dot{y} - \lambda^2 n^2 Ax &= -A\ddot{\theta} - C\lambda\mu\dot{\psi}, \\ A\ddot{y} - C\lambda\mu\dot{x} - \lambda^2 n^2 Ay &= -C\lambda\mu\dot{\theta} - A\ddot{\psi}. \end{aligned} \tag{3.37}$$

Le système d'équations (3.37) est un système de deux équations du second ordre liées, c'est-à-dire un système d'équations différentielles d'ordre quatre.

En procédant aux mêmes estimations dans le système (3.34), on obtient un système d'équations linéaires décrivant le mouvement relatif de la fusée:

$$A\ddot{x} + \varkappa x = -A\ddot{\theta}, \qquad A\ddot{y} + \varkappa y = -A\ddot{\psi},$$

ou, après avoir effectué le changement (3.30),

$$\ddot{x} + \lambda^2 n^2 x = -\ddot{\theta}, \qquad \ddot{y} + \lambda^2 n^2 y = -\ddot{\psi}. \tag{3.38}$$

Ainsi le système (3.38) est un système de deux équations linéaires du second ordre qui se décompose en deux équations indépendantes. Donc pour étudier le mouvement relatif de la fusée dans l'approximation envisagée, il suffit seulement d'étudier le cas plan.

d) *Etude du système d'équations associé et des fonctions frontières dans le cas de la fusée.* Voyons comment le théorème de Tikhonov s'applique à l'analyse des équations du mouvement d'une fusée. Il suffit pour cela de ne considérer qu'une équation du système (3.38), par exemple la première. Dans cette équation on doit poser $t$ = const, $\theta$ = const, et la mettre sous la forme suivante ($\tau = \lambda t$):

$$\frac{d^2x}{d\tau^2} + n^2(t)\,x = 0. \tag{3.39}$$

L'équation (3.39) admet la solution générale

$$x = x_0 \cos(n\tau + \eta), \tag{3.40}$$

où $x_0$ et $\eta$ sont des constantes arbitraires. Donc, la solution triviale de l'équation associée (3.39) est stable au sens de Liapounov et n'est pas asymptotiquement stable, c'est-à-dire que le théorème de Tikhonov est formellement inapplicable. Mais cette conclusion est la conséquence des simplifications faites: nous avons négligé le moment aérodynamique amortissant. Si nous en avions tenu compte, l'équa-

tion associée serait de la forme

$$\frac{d^2x}{d\tau^2}+2a\frac{dx}{d\tau}+n^2x=0 \qquad (3.39')$$

et sa solution triviale serait asymptotiquement stable, c'est-à-dire que la première hypothèse du théorème de Tikhonov serait réalisée. La condition d'attraction de la racine est automatiquement remplie ici, puisque l'équation associée est linéaire.

Ainsi les équations du mouvement de la fusée forment un système tikhonovien classique et la recherche de la solution se ramène à la construction de fonctions frontières et pour cela il faut résoudre le système (3.38). En utilisant la technique développée au § 2 on peut obtenir l'expression de cette solution sous la forme suivante:

$$x=\frac{C}{\sqrt{n(t)}}\cos\left(\lambda\int_0^t n(s)\,ds+\eta\right)-\frac{\ddot{\theta}}{\lambda^2n^2}, \qquad (3.41)$$

où $C$ et $\eta$ sont des constantes arbitraires. On obtient une formule analogue pour $y$:

$$y=\frac{C_1}{\sqrt{n(t)}}\cos\left(\lambda\int_0^t n(s)\,ds+\eta_1\right)-\frac{\ddot{\psi}}{\lambda^2n^2}. \qquad (3.42)$$

Une fois en possession des solutions (3.41) et (3.42), on peut contruire les fonctions frontières pour les variables $v$, $\theta$ et $\psi$. Considérons l'équation (3.27') par rapport à la variable $\psi$:

$$\frac{d\psi}{dt}=R_3-f(t)\,y=\varphi_2(t)\,y,$$

d'où

$$\psi=\psi_0+C_1\int_0^t\varphi_2(s)\frac{1}{\sqrt{n(s)}}\cos\left(\lambda\int_0^s n(\tau)\,d\tau+\eta_1\right)ds-$$
$$-\frac{1}{\lambda^2}\int_0^t\varphi_2(s)\frac{\ddot{\psi}(s)}{n^2(s)}\,ds. \qquad (3.43)$$

Une intégration par parties du premier terme du second membre de (3.43) nous donne

$$\psi=\psi_0+\frac{\varphi_2(t)}{\lambda n\sqrt{n}}\left[A\sin\lambda\int_0^t n(\tau)\,d\tau+B\cos\lambda\int_0^t n(\tau)\,d\tau\right]+O\left(\frac{1}{\lambda^2}\right). \qquad (3.44)$$

Les autres fonctions inconnues du système (3.27') et (3.24) se calculent exactement de la même façon.

Ainsi le problème général de balistique de la fusée à empennage se ramène à l'intégration numérique des équations de son problème fondamental et au calcul de la correction qu'il est nécessaire de faire pour tenir compte de l'influence du mouvement relatif sur le caractère de la trajectoire du centre de masse (trajectoire qui a été calculée sous l'hypothèse que la fusée n'est pas un solide à six degrés de liberté, mais un point matériel à trois degrés de liberté). On voit à partir de la formule (3.44) que le calcul des corrections fait intervenir des quadratures de fonctions à variation lente et des formules explicites.

e) *Etude de l'équation associée et des fonctions frontières dans le cas d'un projectile tournant.* Ce cas est plus compliqué que le problème du mouvement d'une fusée. Cependant on peut conduire les calculs d'après le même schéma classique.

Le système d'équations (3.37) que nous mettrons sous la forme

$$\begin{aligned} &\ddot{x} + 2\lambda e\mu\dot{y} - \lambda^2 n^2 x = F_1(t), \\ &\ddot{y} - 2\lambda e\mu\dot{x} - \lambda^2 n^2 y = F_2(t), \end{aligned} \tag{3.45}$$

où

$$e = \frac{C}{2A}, \quad F_1(t) = -\ddot{\theta} - 2e\lambda\mu\dot{\psi}, \quad F_2(t) = -\ddot{\psi} - 2e\lambda\mu\dot{\theta}.$$

est un système d'ordre quatre. Ce système peut néanmoins être ramené à une seule équation du second ordre à coefficients complexes par rapport à une fonction complexe de la variable réelle $t$.

Introduisons la nouvelle inconnue

$$\xi = x + iy.$$

Multiplions la deuxième équation du système (3.45) par $i$ et ajoutons à la première. On obtient

$$\ddot{\xi} + 2e\lambda\mu\,(\dot{y} - i\dot{x}) - \lambda^2 n^2 \xi = F(t).$$

où

$$F(t) = F_1(t) + iF_2(t).$$

Or $-i\dot{x} - \dot{y} = -i\dot{\xi}$, donc le système (3.45) se ramène finalement à une équation du second ordre

$$\ddot{\xi} - i2e\lambda\mu\dot{\xi} - \lambda^2 n^2 \xi = F(t). \tag{3.46}$$

L'équation (3.46) est une équation linéaire à coefficients variables contenant un grand paramètre. On peut étudier cette équation par les méthodes classiques développées au § 2. Mais avant de passer au calcul des fonctions frontières, élucidons les conditions d'applicabilité du théorème de Tikhonov.

L'équation associée s'écrit ici

$$\frac{d^2\xi}{d\tau^2} - i2e\mu\frac{d\xi}{d\tau} - n^2\xi = F(t),$$

où $\mu$, $F$ et $n$ ne dépendent pas de $\tau$. En introduisant la nouvelle variable $\hat{\xi} = \xi + \frac{F(t)}{n^2(t)}$, on obtient

$$\frac{d^2\hat{\xi}}{d\tau^2} - 2ie\mu\frac{d\hat{\xi}}{d\tau} - n^2\hat{\xi} = 0. \tag{3.47}$$

L'équation (3.47) étant une équation à coefficients constants (par rapport à $\tau$), on peut chercher sa solution sous la forme

$$\hat{\xi} = Ce^{\eta\tau},$$

où $\eta$ est racine de l'équation caractéristique

$$\eta^2 - 2ie\mu\eta - n^2 = 0,$$

soit

$$\eta = ie\mu \pm \sqrt{(n^2 - e^2\mu^2)}. \tag{3.48}$$

L'égalité (3.48) nous dit que si la condition

$$e^2\mu^2 \geqslant n^2 \tag{3.49}$$

n'est pas réalisée, une racine de l'équation caractéristique sera à partie réelle strictement positive et le mouvement du projectile sera instable. La condition (3.49) coïncide, aux notations près, avec la condition de stabilité dans la théorie classique du gyroscope. Nous avons donc indiqué en même temps une méthode d'étude de la stabilité du gyroscope.

Si l'on pose $e^2\mu^2 - n^2 = \sigma^2$ et que l'on admette que la condition (3.49) est réalisée, on peut mettre l'expression (3.48) sous la forme

$$\eta_{1,2} = i(e\mu \pm \delta). \tag{3.50}$$

Donc, l'intégrale générale de l'équation (3.47) est de la forme

$$\hat{\xi} = C_1 \exp\{i(e\mu + \sigma)\tau\} + C_2 \exp\{i(e\mu - \sigma)\tau\}. \tag{3.51}$$

De là il s'ensuit que la solution triviale de l'équation (3.47) est stable au sens de Liapounov, mais n'est pas asymptotiquement stable. Donc, la première hypothèse du théorème de Tikhonov n'est pas remplie.

La stabilité de la solution triviale de l'équation associée qui décrit le mouvement du projectile a été acquise sous l'hypothèse que la résistance de l'air était négligeable. Nous sommes arrivés à la même condition dans le mouvement de la fusée. Mais la situation est fondamentalement différente ici.

Si l'on tient compte aussi des forces dissipatives dans le problème de la fusée, on arrive à l'équation (3.39′) qui décrit les oscillations amorties d'un pendule mathématique. Donc, dans le problème de la fusée, une prise en compte plus exacte des forces agissant sur la fusée fait que la solution triviale de l'équation associée est non seulement stable mais est aussi asymptotiquement stable.

Si les forces de frottement ne sont pas négligées dans le problème du projectile, alors en vertu de la formule (3.19) sa vitesse angulaire décroîtra exponentiellement et par suite la condition (3.29) sera nécessairement violée à un instant $t = t^*$. Donc, pour tous $\omega < \hat{\omega}$ ceci se traduit par une croissance exponentielle de l'angle de nutation qui fera basculer et dévier le projectile de sa trajectoire.

Ainsi le mouvement du projectile est essentiellement instable et le système d'équations qui le décrit n'est pas tikhonovien. Cependant les conditions de stabilité peuvent être violées pour des valeurs de $t$ qui sont *a fortiori* supérieures au temps que mettra le projectile pour atteindre la cible.

On ne peut donc justifier la validité des procédures d'intégration asymptotique en théorie du mouvement du projectile en réduisant le problème à l'analyse de l'applicabilité du théorème de Tikhonov. Mais dans les méthodes développées dans ce chapitre on ne peut compter que sur l'intuition et l'expérience. Or l'expérience de plus d'un siècle de calculs balistiques prouve le bien-fondé et l'efficacité de ces calculs.

REMARQUE. L'expérience des calculs balistiques et les autres exemples d'analyse basés sur l'utilisation des fonctions frontières attestent de la portée de la technique ici développée. Ce qui laisserait supposer que la condition de stabilité asymptotique dans le théorème de Tikhonov peut être affaiblie. Mais à la connaissance de l'auteur, il n'existe pas de résultats concrets en la matière.

Passons maintenant à la construction des fonctions frontières décrivant le mouvement relatif d'un obus d'artillerie. Considérons l'équation (3.46) et effectuons le changement des variables

$$\xi = \zeta \exp\left\{i\lambda e \int_0^t \mu\, dt\right\}$$

ou, si l'on admet que la vitesse angulaire de la rotation propre est constante,

$$\xi = \zeta e^{i\lambda e\mu t}. \qquad (3.52)$$

Des transformations évidentes nous amènent à l'équation différentielle par rapport à $\zeta$:

$$\ddot{\zeta} + \lambda^2\sigma^2(t)\,\zeta = F(t)\, e^{-i\lambda e\mu t}. \qquad (3.53)$$

En vertu des résultats du § 2, une solution particulière de l'équation (3.53) est

$$\tilde{\zeta} = \frac{F(t)\, e^{-i\lambda e \mu t}}{\lambda^2 \sigma^2(t)}. \tag{3.54}$$

Formons maintenant les expressions asymptotiques des solutions linéairement indépendantes de l'équation

$$\ddot{\zeta} + \lambda^2 \sigma^2(t)\, \zeta = 0. \tag{3.55}$$

Nous avons à plusieurs reprises envisagé l'équation (3.55). Son intégrale générale est de la forme

$$\zeta = \frac{1}{\sqrt[4]{e^2\mu^2 - n^2}} \left( C_1 \exp\left[ i\lambda \int_0^t \sqrt{e^2\mu^2 - n^2}\, dt \right] + \right.$$

$$\left. + C_2 \exp\left[ -i\lambda \int_0^t \sqrt{e^2\mu^2 - n^2}\, dt \right] \right),$$

où $C_1$ et $C_2$ sont des constantes arbitraires. En revenant à la variable $\xi$, on obtient la solution générale de l'équation (3.46)

$$\xi = \frac{C_1}{\sqrt[4]{e^2\mu^2 - n^2}} \exp i\lambda \left\{ e\mu t + \int_0^t \sqrt{e^2\mu^2 - n^2}\, dt \right\} +$$

$$+ \frac{C_2}{\sqrt[4]{e^2\mu^2 - n^2}} \exp i\lambda \left\{ e\mu t - \int_0^t \sqrt{e^2\mu^2 - n^2}\, dt \right\} + \frac{F(t)}{\lambda^2 (e^2\mu^2 - n^2)}. \tag{3.56}$$

De la formule (3.56) on déduit immédiatement la condition nécessaire de stabilité

$$e^2\mu^2 \geqslant n^2 \tag{3.57}$$

pour tout $t \in [0, T]$. En effet, si la condition (3.57) n'est pas remplie, alors une solution particulière renfermera nécessairement l'exponentielle à une puissance strictement positive et la théorie considérée n'a plus de sens : le système étudié cesse d'être stable.

Les constantes $C_1$ et $C_2$ de la formule (3.56) sont des nombres complexes

$$C_1 = C_{11} + iC_{12}, \qquad C_2 = C_{21} + iC_{22}. \tag{3.58}$$

L'équation primitive est d'ordre quatre, donc les constantes arbitraires sont au nombre de quatre : $C_{11}$, $C_{12}$, $C_{21}$, $C_{22}$.

Revenons maintenant aux variables réelles en utilisant les notations (3.58) et aussi le fait que $\xi = x + iy$, $F = F_1 + iF_2$. En portant ces expressions dans (3.56), en se servant des formules d'Euler et en comparant les parties réelles et imaginaires de (3.56) on ob-

tient en définitive

$$x = \frac{F_1}{\lambda^2 (e^2\mu^2 - n^2)} + \frac{C_{11}}{\sqrt[4]{e^2\mu^2 - n^2}} \cos\lambda \left[ e\mu t + \int_0^t \sqrt{e^2\mu^2 - n^2}\, dt \right] - \frac{C_{12}}{\sqrt[4]{e^2\mu^2 - n^2}} \sin\lambda \left[ e\mu t + \int_0^t \sqrt{e^2\mu^2 - n^2}\, dt \right] + \frac{C_{21}}{\sqrt[4]{e^2\mu^2 - n^2}} \cos\lambda \left[ e\mu t - \int_0^t \sqrt{e^2\mu^2 - n^2}\, dt \right] - \frac{C_{22}}{\sqrt[4]{e^2\mu^2 - n^2}} \sin\lambda \left[ e\mu t - \int_0^t \sqrt{e^2\mu^2 - n^2}\, dt \right], \quad (3.59)$$

$$y = \frac{F_2}{\lambda^2 (e^2\mu^2 - n^2)} + \frac{C_{11}}{\sqrt[4]{e^2\mu^2 - n^2}} \sin\lambda \left[ e\mu t + \int_0^t \sqrt{e^2\mu^2 - n^2}\, dt \right] + \frac{C_{12}}{\sqrt[4]{e^2\mu^2 - n^2}} \cos\lambda \left[ e\mu t + \int_0^t \sqrt{e^2\mu^2 - n^2}\, dt \right] + \frac{C_{21}}{\sqrt[4]{e^2\mu^2 - n^2}} \sin\lambda \left[ e\mu t - \int_0^t \sqrt{e^2\mu^2 - n^2}\, dt \right] + \frac{C_{22}}{\sqrt[4]{e^2\mu^2 - n^2}} \cos\lambda \left[ e\mu t - \int_0^t \sqrt{e^2\mu^2 - n^2}\, dt \right].$$

Les formules (3.59) peuvent encore s'écrire

$$x = \tilde{x} + x_1 + x_2, \qquad y = \tilde{y} + y_1 + y_2,$$

où $\tilde{x}$ et $\tilde{y}$ sont des solutions particulières:

$$\tilde{x} = \frac{F_1(t)}{\lambda^2 (e^2\mu^2 - n^2)}, \quad \tilde{y} = \frac{F_2(t)}{\lambda^2 (e^2\mu^2 - n^2)}, \qquad (3.60)$$

et $x_i$ et $y_i$ sont définies par les formules

$$x_1 = \frac{A}{\sqrt[4]{e^2\mu^2 - n^2}} \cos\lambda \left[ e\mu t + \int_0^t \sqrt{e^2\mu^2 - n^2}\, dt + \varphi \right],$$

$$y_1 = \frac{A}{\sqrt[4]{e^2\mu^2 - n^2}} \sin\lambda \left[ e\mu t + \int_0^t \sqrt{e^2\mu^2 - n^2}\, dt + \varphi \right], \qquad (3.61)$$

$$x_2 = \frac{B}{\sqrt[4]{e^2\mu^2 - n^2}} \cos\lambda \left[ e\mu t - \int_0^t \sqrt{e^2\mu^2 - n^2}\, dt + \psi \right],$$

$$y_2 = \frac{B}{\sqrt[4]{e^2\mu^2 - n^2}} \sin\lambda \left[ e\mu t - \int_0^t \sqrt{e^2\mu^2 - n^2}\, dt + \psi \right], \qquad (3.62)$$

où $A$, $B$, $\varphi$ et $\psi$ sont de nouvelles constantes arbitraires reliées aux constantes $C_{ij}$ par les formules

$$C_{11} = A\cos\varphi, \qquad C_{12} = A\sin\varphi,$$
$$C_{21} = B\cos\psi, \qquad C_{22} = B\sin\psi.$$

Ces formules montrent que la tête de l'obus (plus exactement, le point d'intersection du vecteur unitaire $\xi^0$ avec le plan perpendiculaire à la vitesse de centre de masse) décrit une courbe qui est la superposition de deux mouvements de rotation. Le premier a lieu suivant un cercle de rayon $\frac{A}{\sqrt[4]{e^2\mu^2-n^2}}$ et sa vitesse angulaire instantanée est $\omega_1 = \lambda\,[e\mu + \sqrt{e^2\mu^2 - n^2}]$. Ce mouvement s'appelle *précession rapide*. Ce mouvement a lieu autour d'un point qui se déplace à son tour suivant un cercle de rayon $\frac{B}{\sqrt[4]{e^2\mu^2-n^2}}$ à une vitesse angulaire instantanée $\omega_2 = \lambda\,[e\mu - \sqrt{e^2\mu^2 - n^2}]$. Ce mouvement s'appelle *précession lente*. Il s'effectue autour d'un point de coordonnées $\tilde{x}$ et $\tilde{y}$.

Le mouvement d'un obus d'artillerie est bien plus compliqué que celui d'une fusée. Le mouvement relatif d'une fusée est identique aux oscillations d'un pendule sphérique. Malgré la grande complexité ou, plus exactement, la lourdeur des formules décrivant les fonctions frontières dans le cas d'un obus, le calcul des autres éléments de la trajectoire ne soulève pratiquement pas de difficultés. Ainsi, les corrections de l'angle d'écart du vecteur vitesse par rapport au plan de tir se calculent à l'aide de l'équation

$$\frac{d\psi}{dt} = \frac{Q_2(x, y, X)}{mv} \approx \frac{q}{v}\,y = ky,$$

d'où

$$\psi = \psi_0 + \int_0^t k(t)\,[\tilde{y} + y_1 + y_2]\,dt, \tag{3.63}$$

où $\tilde{y}$, $y_1$ et $y_2$ sont définies par les formules (3.60), (3.61) et (3.62). En intégrant (3.63) par parties on obtient sans peine des formules identiques à (3.44).

* * *

Dans ce chapitre nous avons étudié une importante classe de problèmes dans lesquels les systèmes mettaient en jeu des variables rapides et des variables lentes. Si ces systèmes satisfont les hypothèses du théorème de Tikhonov, alors les méthodes de ce chapitre nous permettent de les analyser assez efficacement. Le système générateur

correspondant à $\varepsilon = 0$ ne contient pas de fonction dépendant d'une variable rapide et son étude par les méthodes numériques est relativement simple. Le système générateur décrit, pour ainsi dire, le fond sur lequel se déroule le processus, c'est-à-dire sa composante lentement variable. Cette information est en principe presque toujours insuffisante. Bien plus, il arrive parfois que l'analyste s'intéresse précisément à la composante rapidement variable. Par exemple, un problème important en économie est celui des variations d'indices divers sur un fond lentement variable. De nombreux problèmes d'écologie portent sur des explosions temporaires de natalité ou de mortalité de certaines populations, etc. Ces phénomènes impliquent l'étude des fonctions frontières et l'appareil développé ici est utile dans de nombreux cas analogues.

Le schéma de simplification des systèmes d'équations (et la construction d'algorithmes rapides sur sa base) développé dans ce chapitre est une méthode efficace d'analyse préliminaire du problème, une étape qui doit nécessairement précéder toute simulation.

## CHAPITRE VI

# LES MÉTHODES DE LA THÉORIE DES PERTURBATIONS DANS LES PROBLÈMES DE COMMANDE OPTIMALE

## § 1. Schémas élémentaires de théorie des perturbations

a) *Remarques préliminaires.* La simulation du processus étudié par un modèle simple est l'un des problèmes majeurs de l'analyse des systèmes. Nous avons déjà évoqué ce problème et nous reviendrons encore sur les questions soulevées par le passage à des modèles plus simples et à la construction d'algorithmes rapides sur leur base.

Ce problème est à multiples facettes. A. Krylov, l'un des plus grands ingénieurs russes, le père de la première école soviétique d'analyse numérique, se plaisait à répéter une phrase qui se résume ainsi: « Un défaut de précision est une erreur. Un excès de précision, une demi-erreur ». Donc, la précision des calculs doit correspondre à la précision de l'information utilisée, aux possibilités de réalisation de calculs efficaces et surtout aux besoins de la pratique.

Une autre facette du problème est dans les algorithmes rapides qui permettent de procéder à une ébauche du projet, à l'estimation et au choix de l'ensemble des solutions permises, à un devis du projet, etc.

Un autre groupe de questions est lié à la crédibilité des résultats obtenus à l'aide de modèles simples. Les deux chapitres précédents étaient consacrés au développement de l'appareil nécessaire à la résolution de ces problèmes. Dans ces chapitres on a étudié diverses questions liées au comportement des solutions des équations différentielles ordinaires en fonction des variations de certains petits paramètres. Mais jusqu'ici il était question de problèmes d'analyse de systèmes non gouvernés contenant des paramètres, alors que les problèmes de dépendance des solutions par rapport aux paramètres sont particulièrement importants surtout en théorie de la commande: le remplacement d'un modèle d'optimisation par un autre, plus simple, peut réduire de plusieurs fois le volume des calculs.

Malheureusement, il n'est pas toujours possible d'appliquer directement les méthodes d'analyse des propriétés asymptotiques des solutions aux problèmes de commande optimale. Ces problèmes font apparaître de nouvelles difficultés et singularités qui nécessitent une adaptation de l'appareil en vigueur à la nouvelle classe de problèmes.

Dans ce chapitre, nous étudions quelques exemples illustrant ces singularités, en concentrant notre attention sur l'élaboration de la

22—1523

théorie des perturbations, c'est-à-dire aux procédures qui permettent de préciser la solution correspondant aux valeurs nulles des paramètres.

Nous commençons par étudier uniquement les systèmes gouvernés pour lesquels la commande correspondant aux valeurs nulles du paramètre est supposée connue.

Considérons le problème d'optimisation

$$f(x, \varepsilon) \Rightarrow \min \tag{1.1}$$

et désignons par $x_*(\varepsilon)$ sa solution. On distinguera deux cas: le vecteur $x$ n'est soumis à aucune condition (ou est soumis à des liaisons, c'est-à-dire des conditions de type égalités) ; le vecteur $x$ est soumis à des contraintes (c'est-à-dire à des conditions de type inégalités). Traitons le premier cas. Supposons pour fixer les idées que le vecteur $x$ n'est pas borné ou qu'il appartient à un ensemble ouvert. Le vecteur $x_*$ doit alors être solution de l'équation

$$\left.\frac{\partial f(x, \varepsilon)}{\partial x}\right|_{x=x_*} = \varphi(x_*, \varepsilon) = 0. \tag{1.2}$$

Cherchons la solution de l'équation (1.2) sous forme de la série

$$x_*(\varepsilon) = x_*(0) + \varepsilon x_1 + \varepsilon^2 x_2 + \ldots$$

Comme $\varphi(x_*(0), 0) = 0$, le vecteur $x_1$ sera défini par la formule

$$x_1 = -\left\{\frac{\partial\varphi}{\partial x}(x_*(0), 0)\right\}^{-1} \frac{\partial\varphi}{\partial\varepsilon}(x_*(0), 0). \tag{1.3}$$

On remarquera que l'expression $x_*(0) + \varepsilon x_1$ coïncide aux infiniment petits d'ordre deux avec la formule de la méthode de Newton si l'on prend $x_*(0)$ pour première approximation. Prouvons-le. Posons $x = x_*(0) + \delta x$ et récrivons l'équation (1.2) en négligeant les infiniment petits d'ordre $O[(\delta x)^2]$:

$$\varphi(x, \varepsilon) = \varphi(x_*(0), \varepsilon) + \frac{\partial\varphi}{\partial x}(x_*(0), \varepsilon)\,\delta x = 0,$$

d'où

$$\delta x = -\left\{\frac{\partial\varphi}{\partial x}(x_*(0), \varepsilon)\right\}^{-1} \varphi(x_*(0), \varepsilon).$$

Mais

$$\frac{\partial\varphi}{\partial x}(x_*(0), \varepsilon) = \frac{\partial\varphi}{\partial x}(x_*(0), 0) + O(\varepsilon)$$

et

$$\varphi(x_*(0), \varepsilon) = \varphi(x_*(0), 0) + O(\varepsilon),$$

d'où il s'ensuit que

$$x_*(0) + \delta x = x_*(0) + \varepsilon x_1 + O(\varepsilon^2).$$

Donc, dans ce cas (régulier) la théorie des perturbations peut être bâtie par la méthode classique moyennant un développement de la solution en une série du paramètre $\varepsilon$. Si de plus la fonction $\varphi(x, \varepsilon)$ est une fonction analytique du paramètre $\varepsilon$, alors les séries de la forme

$$x = x_* + \varepsilon x_1 + \varepsilon^2 x_2 + \ldots$$

convergent pour des valeurs assez petites de $\varepsilon$. Donc, aux petites valeurs de $\varepsilon$ correspondent de petites variations de la solution $x(\varepsilon)$ du problème (1.1).

La situation est complètement différente lorsque le vecteur $x(\varepsilon)$ est soumis à des contraintes: les raisonnements précédents ne sont plus valables. En effet, le vecteur $x = x_*(0) + \varepsilon x_1$, où $x_1$ se calcule par la formule (1.3), peut se retrouver à l'extérieur du domaine vérifiant les contraintes, donc pour faire usage de la méthode du petit paramètre il faut recourir dans ce cas à une technique spéciale (par exemple, introduire des fonctions de pénalisation). Les problèmes de commande optimale soulèvent de plus grosses difficultés. On connaît quelques classes seulement de problèmes de commande contenant des paramètres qui sont justiciables de méthodes régulières de théorie des perturbations. Commençons leur exposé par l'analyse de deux exemples élémentaires.

b) *Systèmes faiblement gouvernés.* On conviendra d'entendre par de tels systèmes des systèmes qui cessent d'être commandés pour une valeur nulle du paramètre. L'archétype de ces systèmes est

$$\dot{x} = f^{(0)}(x, t) + \varepsilon f^{(1)}(x, t, u), \tag{1.4}$$

où la commande $u$ et la trajectoire de phase $x$ sont soumises aux conditions

$$u \in U, \tag{1.5}$$

$$x(0) = x_0, \tag{1.5'}$$

où $\varepsilon$ est un petit paramètre et $U$ un ensemble borné.

Posons le problème de Lagrange pour ce système: trouver des fonctions $u(t)$ et $x(t)$ satisfaisant les conditions (1.4), (1.5), (1.5′) et réalisant le minimum de la fonctionnelle

$$J(u) = \int_0^T F(x, u)\, dt = \int_0^T \{F^{(0)}(x) + \varepsilon F^{(1)}(x, u)\}\, dt, \tag{1.6}$$

où $T$ est un instant fixe. Cherchons la trajectoire de phase de ce problème sous la forme

$$x = x^{(0)} + \varepsilon x^{(1)}.$$

En négligeant les termes d'ordre $O(\varepsilon)$ dans l'équation (1.4), on obtient le système d'équations suivant

$$\dot{x}^{(0)} = f^{(0)}(x^{(0)}, t),$$

qui ne contient pas la commande et qui définit une seule trajectoire vérifiant les conditions initiales (1.5'). L'équation pour $x^{(1)}$ qui se déduit à partir de (1.4) par élimination des quantités d'ordre $O(\varepsilon^2)$ est de la forme

$$\dot{x}^{(1)} = f_x^{(0)} x^{(1)} + f^{(1)}(x^{(0)}, u, t); \tag{1.7}$$

cette équation contient la commande $u$ qui doit être déterminée à partir de la condition de minimum de la fonctionnelle (1.6).

Transformons la fonctionnelle en n'en gardant que les termes linéaires en $\varepsilon$:

$$J = J^{(0)} + \varepsilon J^{(1)} = \int_0^T F^{(0)}(x^{(0)})\,dt + \varepsilon \int_0^T \{F_x^{(0)} x^{(1)} + F^{(1)}(x^{(0)}, u)\}\,dt. \tag{1.8}$$

Ainsi, la recherche de la commande $u \in U$ et de la « correction » de la trajectoire $x^{(1)}(t)$ s'est ramenée à celle du minimum de la fonctionnelle (1.8) sous la condition (1.7). Par ailleurs, la fonction $x^{(1)}(t)$ doit satisfaire les conditions initiales nulles:

$$x^{(1)}(0) = 0.$$

Voyons ce problème de plus près. Composons le hamiltonien

$$H(\psi, f_x^{(0)} x^{(1)} + f^{(1)}(x^{(0)}, u, t)) - F_x^{(0)} x^{(1)} - F^{(1)}(x^{(0)}, u), \tag{1.9}$$

où la fonction $\psi$ est solution de l'équation

$$\dot{\psi} = -\frac{\partial H}{\partial x^{(1)}} = -\psi f_x^{(0)} + F_x^{(0)}. \tag{1.10}$$

Le hamiltonien dépendant linéairement de la variable de phase $x^{(1)}$, l'équation (1.10) ne contient pas $x^{(1)}$ et peut être intégrée indépendamment de l'équation (1.7).

La spécificité du problème considéré est due dans une large mesure à la nature des conditions portant sur l'extrémité droite de la trajectoire. Attardons-nous uniquement sur deux cas extrêmes: l'extrémité droite est: 1° libre; 2° fixe. Dans le premier cas, la condition de transversalité nous donne

$$\psi(T) = 0. \tag{1.11}$$

Dans le second cas

$$x(T) = x_T, \tag{1.12}$$

où $x_T$ est un vecteur donné.

Dans le premier cas, la résolution du problème de Cauchy (1.10), (1.11) nous donne une seule fonction du temps $\psi(t)$. Donc, à partir du principe du maximum $H \Rightarrow \max$, on déduit le problème suivant de commande optimale:

$$(\psi, f^{(1)}(x^{(0)} u, t)) - F^{(1)}(x^{(0)}, u) \Rightarrow \max_{u \in U}. \qquad (1.13)$$

La variable de phase $x^{(1)}$ ne figure pas dans l'expression (1.13), c'est-à-dire que dans ce cas on trouve la solution du problème de synthèse, puisque la commande $u(\cdot)$ se détermine indépendamment des valeurs de la fonction vectorielle $x^{(1)}(t)$. Après avoir acquis la commande $u = u(t)$ à partir de la condition (1.13), on définit la trajectoire de phase $x^{(1)}(t)$ à l'aide de l'équation (1.7).

La linéarisation du problème et le passage au système (1.7) peuvent inopinément prolonger les calculs. Dans ce cas, la méthode des approximations successives peut s'avérer meilleure à l'usage. La marche à suivre est la suivante. Au premier pas on résout le problème de Cauchy suivant:

$$\dot{x}^{(0)} = f^{(0)}(x^{(0)}, t); \quad x(0) = x_0. \qquad (1.14)$$

On calcule, puis on mémorise la quantité $x^{(0)}(T) = x_T^{(0)}$. La trajectoire $x^{(0)}(t)$ n'est pas mémorisée. On résout ensuite le problème de Cauchy suivant

$$x^{(0)}(T) = x_T^{(0)}, \quad \psi^{(0)}(T) = 0$$

pour les équations (1.10) et (1.14) et on déduit la commande $u^{(0)}(t)$ à partir de (1.13). Au pas suivant on résout le problème de Cauchy:

$$x^{(1)} = f^{(0)}(x^{(1)}, t) + f^{(1)}(x^{(1)}, t, u^{(0)}; \; x^{(1)}(0) = x_0, \qquad (1.15)$$

et ainsi de suite.

Un élément d'analyse très important est l'estimation de l'erreur, c'est-à-dire l'estimation de l'écart entre la valeur minimale $J^*$ de la fonctionnelle et la valeur $\hat{J}$ de cette fonctionnelle obtenue par la méthode d'approximation décrite plus haut. A noter que l'écart entre les diverses trajectoires de phase (cf. (1.4)) est de l'ordre de $O(\varepsilon)$ quelles que soient les commandes admissibles $u \in U$. Donc, on a la même estimation pour la fonctionnelle $J(u)$:

$$J^* = \hat{J} + O(\varepsilon). \qquad (1.16)$$

Mais comme l'a montré A. Lioubouchine [50] *), l'estimation (1.16) peut être nettement améliorée. Si pour des $\varepsilon$ assez petits, la solution de l'équation (1.4) est uniformément bornée pour tous $u \in U$, liors

$$J^* = \hat{J} + O(\varepsilon^2).$$

*) Le cas général du problème de Mayer pour les systèmes faiblement gouvernés a été traité par F. Tchernoousko [66].

REMARQUE. On voit sans peine que l'algorithme développé peut être traité comme une modification de la méthode de Krylov-Tchernoousko pour le cas où nous avons une « bonne » première approximation. En effet, considérons le problème de minimisation de la fonctionnelle $J$ sous la condition

$$\dot{x} = f(x, t, u).$$

Supposons par ailleurs que nous avons une commande « proche » de la commande optimale $\hat{u}(t)$. Posons

$$u = \hat{u} + \varepsilon v,$$

et faisons le changement

$$f = f^{(0)}(x, t) + \varepsilon f^{(1)}(x, t, v)$$

où $f^{(0)} = f(x, \hat{u}, t)$, $f_1 = f_u v$.

Voyons maintenant le deuxième cas, c'est-à-dire la condition aux limites (1.12). La situation est bien plus complexe ici, car la condition aux limites pour l'équation des impulsions (1.10) fait défaut. Néanmoins la détermination de $x^{(1)}(t)$ et de la commande $u(t)$ est plus simple que le problème initial, car l'équation pour les impulsions ne contient pas de variable de phase et, par suite, la commande que nous chercherons à l'aide de la condition (1.13) ne dépendra pas non plus explicitement de la variable de phase. Ces circonstances simplifient de façon notoire la procédure de détermination des valeurs aux bornes des impulsions $\psi(t)$ (ou $\psi(0)$) réalisant la condition (1.12).

Les autres problèmes d'optimisation se traitent de façon analogue. Considérons par exemple le problème de Mayer avec la fonctionnelle linéaire

$$J = (c, x(T)).$$

On sait que tout problème de Mayer peut être ramené à cette forme. Pour construire la théorie des perturbations on pose de nouveau

$$x = x^{(0)} + \varepsilon x^{(1)}; \tag{1.17}$$

on est ainsi conduit à minimiser la fonctionnelle $J_1 = (c, x^{(1)}(T))$ sous la condition

$$\dot{x}^{(1)} = f_x^{(0)} x^{(1)} + f^{(1)}(x^{(0)}, u).$$

On déterminera la commande à partir de la condition

$$(\psi, f^{(1)}(x^{(0)}, u)) \Rightarrow \max_{u \in U},$$

où

$$\dot{\psi} = -\psi f_x^{(0)}, \qquad \psi(T) = -c.$$

Donc, ici aussi on trouve la commande indépendamment du vecteur $x^{(1)}$.

c) *Autre approche du problème posé.* En étudiant les systèmes dynamiques faiblement gouvernés de la forme

$$\dot{x} = f(x, \varepsilon u), \qquad (1.18)$$

$$J = J(x, \varepsilon u), \qquad (1.19)$$

on s'est servi de la procédure suivante de calcul de la trajectoire et de la commande: on a admis que la représentation (1.17) était valable, on l'a portée dans l'équation (1.18) et dans la fonctionnelle (1.19), puis on n'a conservé que les termes linéaires en $x^{(1)}$. On a ensuite composé un nouveau problème variationnel pour trouver la commande $u(t)$ et la « correction » de la trajectoire de phase $x^{(1)}(t)$. Mais on peut procéder autrement.

Considérons le système (1.18), (1.19), où la commande et la trajectoire de phase sont soumises aux conditions (1.5), (1.5′) et où la fonctionnelle est de la forme

$$J = F(x(T)). \qquad (1.20)$$

Composons le hamiltonien

$$H = (\psi, f(x, \varepsilon u)) \qquad (1.21)$$

et l'équation pour les impulsions

$$\dot{\psi} = -\frac{\partial H}{\partial x} = -f_x^* \psi. \qquad (1.22)$$

Cherchons la solution sous la forme

$$x = x^{(0)} + \varepsilon x^{(1)} + \varepsilon^2 x^{(2)} + \ldots, \quad u = u^{(0)} + \varepsilon u^{(1)} + \varepsilon^2 u^{(2)} + \ldots,$$

$$\psi = \psi^{(0)} + \varepsilon \psi^{(1)} + \varepsilon^2 \psi^{(2)} + \ldots$$

En portant ces séries dans les équations (1.18), (1.22) et en identifiant les coefficients des mêmes puissances de $\varepsilon$, on obtient des équations qui sont vérifiées par les fonctions $x^{(i)}$ et $\psi^{(i)}$. Pour les fonctions $x^{(0)}$ et $\psi^{(0)}$ on obtient notamment les équations suivantes:

$$\dot{x}^{(0)} = f(x^{(0)}, 0), \qquad (1.23)$$

$$\dot{\psi}^{(0)} = -f_x^*(x^{(0)}, 0)\, \psi^{(0)}. \qquad (1.24)$$

Faisons de même avec l'expression de la fonctionnelle

$$J = F(x^{(0)}(T)) + \varepsilon J_1 + \varepsilon^2 J_2 + \ldots,$$

où

$$J_1 = (c, x^{(1)}), \quad c = \left(\frac{\partial F}{\partial x}\right)_{x = x^{(0)}(T)}, \quad \text{et ainsi de suite.}$$

La fonction $\psi$ satisfait les conditions de transversalité suivantes:

$$\psi(T) = -\left(\frac{\partial F}{\partial x}\right)_{t=T}.$$

Comme

$$\left(\frac{\partial F}{\partial x}\right)_{t=T} = \left(\frac{\partial F}{\partial x}\right)_{\substack{x=x^{(0)}\\ t=T}} + \varepsilon\left(\frac{\partial^2 F}{\partial x^2}\right)_{\substack{x=x^{(0)}\\ t=T}} x^{(1)} + \ldots,$$

on obtient pour $\psi^{(0)}(T)$ la condition aux limites suivantes:

$$\psi^{(0)}(T) = -\left(\frac{\partial F}{\partial x}\right)_{\substack{x=x^{*}\\ t=T}} = -c.$$

En vertu de (1.23) l'équation pour $x^{(0)}(t)$ ne contient pas la commande et elle définit une seule fonction $x^{(0)}(t)$ vérifiant les données de Cauchy. La fonction $\psi^{(0)}(t)$ est définie uniquement par le vecteur $x^{(0)}(t)$. Une fois qu'on connaît $x^{(0)}$ et $\psi^{(0)}$, on déduit la commande $u^{(0)}$ à partir de la condition

$$(\psi^{(0)}, f(x^{(0)}, \varepsilon u^{(0)})) \Rightarrow \max,$$

d'où il s'ensuit que la commande $u^{(0)}$ ne dépend pas de $x^{(1)}$. Les deux approches nous conduisent donc au même résultat en première approximation. Mais ces deux méthodes ne sont pas pour autant équivalentes. Ainsi, par exemple, en composant le système d'équations (1.18), (1.22), on écrit la condition nécessaire (si $U$ est un ensemble ouvert) sous la forme

$$\frac{\partial H}{\partial u} = 0. \qquad (1.25)$$

Si les seconds membres du système (1.18), (1.22) et la condition (1.25) dépendent analytiquement du paramètre $\varepsilon$, alors les conditions du théorème de Poincaré sont remplies et l'on peut représenter la trajectoire de phase et la commande par des séries convergentes de $\varepsilon$. Donc, dans ces conditions, la méthode d'analyse développée permet d'obtenir la solution exacte du problème d'optimisation primitif, tandis que les raisonnements du nº b) ne permettent de calculer que les premières corrections, c'est-à-dire de construire une variante élémentaire de la théorie des perturbations.

d) *Commandes localement optimales.* Les *commandes localement optimales* jouent un grand rôle en théorie de la synthèse et dans les systèmes de commande complexes. Par ce terme on entendra des commandes qui sont déduites à chaque instant donné à partir de la condition de minimum d'une certaine quantité scalaire. Il existe plusieurs types de commandes localement optimales. Nous n'en citerons que deux. Soit donné un système commandé dont le mouvement est régi par l'équation

$$\dot{x} = f(x, u).$$

Désignons la trajectoire de phase programmée par $z = z(t)$. Supposons que pour une raison quelconque le système a quitté la trajectoi-

re programmée et que $x(t_0) \neq z(t_0)$ à un instant $t = t_0$. Nous devons donc utiliser la commande correctrice $u$ pour ramener le système sur la trajectoire programmée. Introduisons la fonction scalaire

$$F(y) = \frac{1}{2}(y, Ry), \tag{1.26}$$

où $R$ est une matrice symétrique définie positive et $y = z - x$. Alors la commande localement optimale sera une commande qui à chaque instant minimisera la dérivée $dF/dt$, c'est-à-dire

$$\frac{dF}{dt} = \left(\frac{dy}{dt}, Ry\right) \Rightarrow \min_{u \in U}. \tag{1.27}$$

La fonction (1.26) définit la «distance» de la position réelle du système à la position programmée. Donc, la signification de la commande définie par la condition (1.27) est évidente: minimiser cette distance à chaque instant donné.

L'autre exemple de commande localement optimale nous est fourni par la théorie de la commande terminale. Supposons que l'objectif de la commande est d'atteindre un état final (ou terminal) bien défini

$$x(T) = x_T.$$

Introduisons une fonction caractérisant la « distance » au but de la commande:

$$F(x) = \frac{1}{2}[(x - x_T), R(x - x_T)] \tag{1.28}$$

et déduisons la commande à partir de la condition

$$\frac{dF}{dt} \Rightarrow \min. \tag{1.29}$$

(A noter que ce problème avec une fonction objectif de la forme (1.28) se rencontre aussi lors de la construction d'algorithmes de la théorie de la rapidité.)

La commande localement optimale définie à partir de la condition (1.29) et la commande optimale du problème

$$t = T, \quad F(x) \Rightarrow \min,$$

où $F(x)$ est donnée par la formule (1.28) et où la variable de phase obéit à l'équation

$$\dot{x} = f(x, u), \tag{1.30}$$

n'ont rien de commun entre elles dans le cas général. Toutefois, dans le cas du problème de Mayer relatif aux systèmes de la forme (1.18), que nous avons convenu d'appeler faiblement gouvernés, les commandes localement optimales sont des asymptotes pour les com-

mandes optimales et peuvent être utilisées pour les représenter approximativement. Ce fait a visiblement été mis en évidence « expérimentalement » pour la première fois par V. Lébédev [49]. En étudiant les problèmes complexes d'optimisation des trajectoires des engins cosmiques équipés d'un moteur de faible poussée (voile solaire ou moteur électronucléaire), il a découvert que la solution exacte, qui réclame plusieurs heures de temps machine, peut être bien approchée par une commande localement optimale. Le calcul de la trajectoire d'un engin guidé à l'aide d'une commande localement optimale a nécessité quelques secondes de temps machine et la précision fournie satisfaisant parfaitement aux besoins de la pratique et correspondant à l'information initiale.

Plus tard, ce fait a été rigoureusement démontré par F. Tchernoousko [66]. Reprenons, en suivant F. Tchernoousko, les raisonnements qui prouvent que les commandes localement optimales approchent les commandes optimales dans les cas où ces dernières sont petites.

Soit donc un mouvement commandé régi par l'équation (1.18). Supposons que la fonction $f$ est continûment dérivable par rapport aux deux variables, mettons l'équation (1.18) sous la forme

$$\dot{x} = f(x, 0) + \varepsilon B u + O(\varepsilon^2), \tag{1.31}$$

où $B = \left(\frac{\partial f(x, y)}{\partial y}\right)_{y=0}$, et négligeons les infiniment petits du deuxième ordre. Considérons maintenant l'équation

$$\dot{x} = f(x, 0),$$

qui décrit un mouvement non commandé. On admettra que son intégrale générale est de la forme

$$x = \Psi(t, c), \tag{1.32}$$

où $c$ est le vecteur des constantes arbitraires. L'expression (1.32) peut être traitée comme une formule de changement des variables. En passant de la variable $x$ à la variable $c$ dans l'équation (1.31), on obtient une nouvelle équation. Calculons pour cela

$$\frac{d\Psi}{dt} = \frac{\partial \Psi}{\partial t} + \frac{\partial \Psi}{\partial c} \frac{dc}{dt}.$$

Mais $\Psi(t, c)$ est l'intégrale générale de l'équation $\dot{x} = f(x, 0)$, c'est-à-dire que la fonction $\Psi$ vérifie identiquement cette équation par rapport à $c$:

$$\frac{\partial \Psi}{\partial t} = f(\Psi, 0).$$

Donc

$$\frac{dx}{dt} = \frac{d\Psi}{dt} = f(\Psi, 0) + \frac{\partial \Psi}{\partial c} \frac{dc}{dt}.$$

Mais par ailleurs

$$\frac{dx}{dt} = f(x, 0) + \varepsilon Bu.$$

Donc, la nouvelle variable $c$ est solution de l'équation suivante:

$$\dot{c} = \varepsilon \left(\frac{\partial \Psi}{\partial c}\right)^{-1} Bu. \qquad (1.33)$$

Ecrivons maintenant la fonctionnelle:

$$F(x(T)) = F(\Psi(T, c(T))) = F^*(c(T)). \qquad (1.34)$$

Nous avons ainsi ramené le problème initial à la minimisation de la fonctionnelle (1.34) sous la condition (1.33) en commettant une erreur de l'ordre de $\varepsilon^2$. Cherchons la solution de ce problème sous la forme

$$c = c_0 + \varepsilon c_1 + O(\varepsilon^2),$$

où $c_0$ est un vecteur constant et $c_1$ est solution de l'équation

$$\dot{c}_1 = \left(\frac{\partial \Psi}{\partial c}\right)^{-1}_{c=c_0} Bu. \qquad (1.35)$$

Transformons la fonctionnelle (1.34):

$$F^*(c(T)) = F^*(c_0) + \left(\varepsilon \left(\frac{dF^*}{dc}\right)_{c=c_0}, c_1(T)\right) + O(\varepsilon^2).$$

Nous sommes ainsi conduits à déterminer une commande réalisant le minimum de la fonctionnelle

$$J = \left(\varepsilon \left(\frac{\partial F^*}{\partial c}\right)_{c=c_0}, c_1(T)\right)$$

sous la condition (1.35). Composons le hamiltonien

$$H = \left(\psi, \left(\frac{\partial \Psi}{\partial c}\right)^{-1}_{c=c_0} Bu\right)$$

et les équations pour les variables adjointes

$$\dot{\psi} = 0, \quad \psi(T) = -\varepsilon \left(\frac{\partial F^*}{\partial c}\right)_{c=c_0}.$$

Il est immédiat que

$$H = -\varepsilon \left(\left(\frac{\partial F^*}{\partial c}\right)_{c=c_0}, \quad \left(\frac{\partial \Psi}{\partial c}\right)^{-1}_{c=c_0} Bu\right).$$

Le principe du maximum nous dit que la commande doit être déduite à partir de la condition

$$\left(\left(\frac{\partial F^*}{\partial c}\right)_{c=c_0}, \left(\frac{\partial \Psi}{\partial c}\right)^{-1}_{c=c_0} Bu\right) \Rightarrow \min_{u \in U}. \qquad (1.36)$$

Construisons maintenant la commande localement optimale pour le système (1.33) en nous souvenant que nous avons composé ce dernier avec une erreur de l'ordre de $\varepsilon^2$.

D'après ce qui précède, la commande localement optimale sera définie à partir de la condition

$$\frac{dF^*}{dt} \Rightarrow \min.$$

Mais

$$\frac{dF^*(c(T))}{dt} = \left(\frac{dF^*}{dc}, \dot{c}\right) = \varepsilon \left(\frac{dF^*}{dc}, \left(\frac{\partial \Psi}{dc}\right)^{-1} Bu\right). \qquad (1.37)$$

En comparant les expressions (1.36) et (1.37), on constate qu'elles ne diffèrent que par le facteur $\varepsilon$.

Ainsi, la commande localement optimale est justiciable de l'estimation

$$J(u^*) = \hat{J} + O(\varepsilon^2),$$

où $\hat{J}$ est la valeur optimale de la fonctionnelle. Donc, la commande localement optimale est d'autant plus proche de l'optimale que la commande correctrice est faible, c'est-à-dire que $\varepsilon$ est petit.

Les commandes localement optimales comptent parmi les plus importants instruments de construction des algorithmes rapides pour la résolution des problèmes de commande optimale. Non seulement elles permettent de construire des commandes proches des optimales pour une vaste classe de problèmes, mais elles peuvent encore servir de premières approximations dans les diverses procédures itératives, car elles sont admissibles. Par ailleurs, elles peuvent être utilisées pour l'organisation des systèmes de rétroaction, car les commandes correctrices sont généralement petites.

Soit à maximiser la fonctionnelle $F(x(T))$, par exemple, le coût d'un produit fini. Supposons que ce processus est régi par l'équation

$$\dot{x} = f(x, t, u),$$

où $u$ désigne comme toujours la commande, en l'occurrence les ressources. Désignons par $\hat{u}(t)$ un procédé traditionnel de son utilisation et par $v(t)$, $v(t) \in G$, les ressources supplémentaires nécessitées par la correction (correction rendue nécessaire par l'apparition de facteurs imprévisibles). Posons par ailleurs

$$u = \hat{u} + v.$$

Supposons qu'à un instant $t = t_0$ on constate que la trajectoire $x(t_0)$ du processus s'écarte de la trajectoire programmée $x = \hat{x}(t_0)$. Le responsable a le choix entre deux lignes d'action. Tout d'abord il

peut utiliser ses ressources $v$ pour ramener le système sur la trajectoire programmée. Mais dans le cas considéré ce comportement n'est certainement pas le plus raisonnable. Il est préférable de choisir la commande correctrice de manière à maximiser la valeur finale de la fonctionnelle en partant de l'état $x(t_0)$ du système. Comme la fonction correctrice $v$ est petite, il n'est pas indispensable, en vertu du théorème qui vient juste d'être établi, de résoudre de nouveau le problème de commande optimale : il suffit d'utiliser une commande localement optimale qui sera donnée par la condition

$$\frac{dF}{dt} \Rightarrow \max \quad \forall t,$$

ou

$$\frac{dF}{dt} f(x, \hat{u} + v) \Rightarrow \max.$$

e) *Utilisation de la méthode de factorisation.* Les variantes de la théorie des perturbations exposées dans les numéros précédents de ce paragraphe sont assez commodes pour résoudre des problèmes à extrémité libre. Mais les calculs se compliquent énormément lorsque les coordonnées de l'extrémité droite de la trajectoire doivent satisfaire des conditions supplémentaires. Dans ces problèmes, il est préférable de modifier les raisonnements ci-dessus.

Au chapitre II, nous avons développé une méthode de résolution approchée des problèmes d'optimisation qui utilisait une procédure permettant de ramener ces problèmes à une suite de problèmes aux limites pour des équations différentielles linéaires. Comme cette classe de problèmes aux limites est justiciable de méthodes de résolution régulières mettant en jeu des schémas de factorisation stables, nous obtenons des méthodes efficaces de résolution numérique des problèmes d'optimisation se ramenant à des problèmes aux limites linéaires. Cette approche peut être étendue aux problèmes de commande contenant de petits paramètres.

Soit à minimiser la fonctionnelle

$$J(u) = \int_0^T F(x, u, \varepsilon)\, dt \tag{1.38}$$

sous les conditions

$$\dot{x} = f(x, u, \varepsilon), \tag{1.39}$$

$$x(0) = x_0, \quad x(T) = x_T. \tag{1.40}$$

Nous n'imposerons pas d'autres conditions aux commandes et aux trajectoires de phase : soit que ces conditions n'existent pas, soit que des fonctions de pénalisation aient été introduites durant le processus de formation de la fonctionnelle $J(u)$ pour les éliminer. La spé-

cificité des conditions aux limites (1.40) n'est pas essentielle. Les raisonnements qui vont suivre se généralisent sans peine aux autres cas de conditions aux limites. Les fonctions $F$ et $f$ seront supposées suffisamment dérivables par rapport à leurs arguments.

Supposons que nous sachions résoudre le problème (1.38), (1.39), (1.40) pour $\varepsilon = 0$. Désignons cette solution par $x^*$ et $u^*$ et posons

$$x = x^* + \varepsilon y, \qquad u = u^* + \varepsilon v. \tag{1.41}$$

En portant les expressions (1.41) dans l'équation (1.39) et en ne retenant que les termes du premier ordre en $\varepsilon$, on obtient l'équation suivante par rapport à $y$ et $v$:

$$\dot{y} = Ay + Bv + f_\varepsilon, \tag{1.42}$$

où $A = \left(\frac{\partial f^{(i)}}{\partial x^{(j)}}\right)_{\substack{x=x^*\\ u=u^*\\ \varepsilon=0}}$, $B = \left(\frac{\partial f^{(i)}}{\partial u^{(j)}}\right)_{\substack{x=x^*\\ u=u^*\\ \varepsilon=0}}$ sont des matrices et $f_\varepsilon = \left(\frac{\partial f^{(i)}}{\partial \varepsilon}\right)_{\substack{x=x^*\\ u=u^*\\ \varepsilon=0}}$ un vecteur.

Transformons la fonctionnelle $J(u)$ d'une autre façon. Portons les expressions (1.41) dans (1.38) et gardons les termes du premier et du deuxième ordre de petitesse. Autrement dit, mettons la fonctionnelle $J(u)$ sous la forme

$$J(u) = J^*(u^*) + \varepsilon \hat{J}(v, \varepsilon) + \tilde{J},$$

où $\tilde{J}$ est l'ensemble des termes ne dépendant pas de $v$. Calculons $\hat{J}$. A cet effet, développons la fonction $F(x, u, \varepsilon)$ en une série de Taylor:

$$F(x, u, \varepsilon) = F(x^* + \varepsilon y, u^* + \varepsilon v, \varepsilon) = F(x^*, u^*, 0) + \\ + \varepsilon\left(\frac{\partial F}{\partial x} y + \frac{\partial F}{\partial u} v + \frac{\partial F}{\partial \varepsilon}\right) + \frac{\varepsilon^2}{2}\left[\left(y, \frac{\partial^2 F}{\partial x^2} y\right) + \left(v, \frac{\partial^2 F}{\partial u^2} v\right) + \right. \\ \left. + \left(y, \frac{\partial^2 F}{\partial u\, \partial x} v\right) + \left(v, \frac{\partial^2 F}{\partial u\, \partial x} y\right) + \frac{\partial^2 F}{\partial x\, \partial \varepsilon} y + \frac{\partial^2 F}{\partial u\, \partial \varepsilon} v + \frac{\partial^2 F}{\partial \varepsilon^2}\right] + O(\varepsilon^3).$$

Vu que les dérivées ont été calculées pour $\varepsilon = 0$, les termes $\partial F/\partial\varepsilon$ et $\partial^2 F/\partial\varepsilon^2$ ne contiennent ni les commandes correctrices $v$, ni les corrections des variables de phase $y$ et ils peuvent être négligés. Le problème se ramène donc à la minimisation de la fonctionnelle

$$\hat{J}(v, \varepsilon) = \int_0^T \{(a, y) + (b, v) + \varepsilon[(y, C_1 y) + (v, C_2 v) + (v, C_3 y)]\}\, dt, \tag{1.43}$$

où $a$ et $b$ sont des vecteurs de composantes respectives

$$a_i = \frac{\partial F}{\partial x^i} + \varepsilon \frac{\partial^2 F}{\partial x^i \, \partial \varepsilon}, \quad b_i = \frac{\partial F}{\partial u^i} + \varepsilon \frac{\partial^2 F}{\partial u^i \, \partial \varepsilon}$$

et

$$C_1 = \frac{1}{2} \left( \frac{\partial^2 F}{\partial x^i \, \partial x^j} \right), \quad C_2 = \frac{1}{2} \left( \frac{\partial^2 F}{\partial u^i \, \partial u^j} \right),$$

$$C_3 = \left( \frac{\partial^2 F}{\partial x^i \, \partial u^j} \right) + \left( \frac{\partial^2 F}{\partial x^i \, \partial u^j} \right)^*$$

des matrices, le symbole ( )* désignant la transposition. Toutes ces quantités sont calculées pour $x = x^*$, $u = u^*$ et $\varepsilon = 0$.

La fonction $x^* (t)$ satisfait les conditions aux limites (1.40), puisque nous avons admis que nous savions résoudre le problème (1.38), (1.39), (1.40) pour $\varepsilon = 0$. La fonction $y (t)$ doit donc vérifier les conditions aux limites nulles

$$y (0) = y (T) = 0. \tag{1.44}$$

La variante de la théorie des perturbations proposée ramène ainsi la recherche des commandes correctrices $v$ et des corrections de la trajectoire de phase $y$ au problème d'optimisation suivant : trouver le minimum de la fonctionnelle (1.43) sous les conditions (1.42), (1.44). C'est un problème de minimisation d'une fonctionnelle quadratique sous des liaisons différentielles linéaires qui est bien étudié. Nous en avons parlé en détail au chapitre II : le principe du maximum ramène ce problème à un problème aux limites pour un système d'équations différentielles linéaires. En effet, composons le hamiltonien du problème (1.42), (1.43), (1.44) :

$$H = (\psi, Ay + Bv + f_\varepsilon) - (a, y) - (b, v) - \varepsilon (y, C_1 y) - \\ - \varepsilon (v, C_2 v) - \varepsilon (v, C_3 y).$$

Le vecteur des impulsions doit vérifier l'équation suivante :

$$\dot{\psi} = -\frac{\partial H}{\partial y} = -A^* \psi + a + \varepsilon \widetilde{C}_1 y + \varepsilon C_3^* v, \tag{1.45}$$

où $\widetilde{C}_1 = C_1 + C_1^*$. Comme la commande n'est soumise à aucune contrainte, la condition nécessaire de maximum du hamiltonien sera

$$\frac{\partial H}{\partial v} = 0$$

ou

$$B^* \psi - b - \varepsilon C_3 y - \varepsilon \widetilde{C}_2 v = 0, \tag{1.46}$$

où $\widetilde{C}_2 = C_2 + C_2^*$. De l'équation (1.46) il s'ensuit que la commande $v$ sera une fonction linéaire de la variable de phase $y$ et de l'impul-

sion $\psi$:

$$v = d + D_1 y + D_2 \psi, \tag{1.47}$$

où $d$, $D_1$, $D_2$ sont des fonctions connues du temps ($d$ est un vecteur, $D_1$ et $D_2$, des matrices). En portant l'expression (1.47) dans les équations (1.42) et (1.45), on obtient un système d'équations différentielles linéaires par rapport à $y$ et $\psi$ de la forme

$$\begin{aligned} \dot{y} &= A_{11} y + A_{12} \psi + f_1, \\ \dot{\psi} &= A_{21} y + A_{22} \psi + f_2. \end{aligned} \tag{1.48}$$

Nous devons résoudre le problème aux limites (1.44) pour ce système d'équations. Au chapitre II, nous avons exposé en détail le schéma de factorisation stable qui permet de ramener le problème (1.44), (1.48) à un problème de Cauchy et, par conséquent, de le résoudre par une méthode régulière.

REMARQUES. 1. Nous avons admis que la solution exacte du problème était connue pour $\varepsilon = 0$. Dans les problèmes pratiques, on peut en général affaiblir cette condition. Pour approximation initiale $x^*$, $u^*$ on peut prendre, par exemple, une trajectoire qui transfère le système non pas en $x_T$, mais en un point $\hat{x}_T$. Au lieu des conditions homogènes (1.44), on aura alors pour la fonction $y$

$$(\hat{x} + \varepsilon y)_{t=T} = x_T,$$

ou

$$y(T) = (x_T - \hat{x}_T)/\varepsilon. \tag{1.49}$$

Cette méthode marche bien si la quantité (1.49) est « assez » petite.

2. Vu que la théorie des perturbations développée est basée sur la linéarisation du problème pour $\varepsilon$ fini (et non pas infiniment petit), il n'est pas évident que l'inégalité suivante

$$J(u^* + \varepsilon v) < J(u^*) \tag{1.50}$$

aura toujours lieu. Donc, la vérification de la condition (1.50) est obligatoire. Si cette condition est violée, il faut alors tenter de remplacer la commande $u^* + \varepsilon v$ par la commande $\hat{u} = u^* + k\varepsilon v$, où $k < 1$. Mais les conditions aux limites risquent d'être violées. Par conséquent, le procédé qui a été proposé pour préciser la commande optimale n'est pas universel (de même d'ailleurs que toute théorie des perturbations).

3. Si $f$ et $F$ sont des fonctions analytiques de leurs variables, alors d'après ce qui précède on peut construire un schéma régulier de détermination de la solution exacte du problème d'optimisation. Ceci étant, la solution approchée trouvée à l'aide de la méthode étudiée dans ce numéro peut être prise pour première approximation (et la solution $x^*$, $u^*$, pour approximation d'ordre zéro). Désignons cette première approximation par $x_1$ et $u_1$, c'est-à-dire que $x_1 = x^* + \varepsilon y$, $u_1 = u^* + \varepsilon v$. Désignons encore la solution exacte du problème d'optimisation par $x^{**}$ et $u^{**}$ et les valeurs correspondantes de la fonctionnelle (1.38) par $J^*$, $J_1$ et $J^{**}$. Alors, en appliquant le théorème de Poincaré, on obtient les estimations suivantes:

$$x^* = x^{**} + O(\varepsilon), \quad u^* = u^{**} + O(\varepsilon), \quad J^* = J^{**} + O(\varepsilon),$$
$$x_1 = x^{**} + O(\varepsilon^2), \quad u_1 = u^{**} + O(\varepsilon^2), \quad J_1 = J^{**} + O(\varepsilon^2).$$

## § 2. Méthodes de moyennisation dans les problèmes de commande optimale

a) *Systèmes commandés à phase tournante.* Nous avons vu au chapitre IV que le domaine d'application des méthodes de moyennisation pour l'analyse des systèmes et notamment des systèmes à éléments oscillants ou tournants était très vaste. Au chapitre IV, nous avons développé en détail des méthodes de calcul des trajectoires de phase pour des systèmes ne contenant pas les commandes.

Dans ce paragraphe, on se propose d'étudier des systèmes commandés ne contenant pas d'éléments oscillants ou tournants. En traitant les problèmes variationnels pour ces systèmes, on constate que le $\pi$-système (ou système d'équations de L. Pontriaguine) obtenu à l'aide du principe du maximum contient des variables aussi bien rapides que lentes. Cette particularité du système pontriaguinien complique énormément la recherche des extrémums, car elle engendre des « vallées ». Donc, la séparation des mouvements lents et des mouvements rapides dans les systèmes commandés est encore plus importante que dans les systèmes non commandés.

Considérons un exemple classique de système commandé à élément oscillant qui est identique aux exemples qui nous ont servi à développer les algorithmes de séparation des mouvements au chapitre IV. Désignons par $z$ $(t)$ le vecteur décrivant la variation lente du « fond » :

$$\dot{z} = \varepsilon Z\,(z, \alpha, u), \tag{2.1}$$

où $\alpha$ est une composante scalaire oscillante obéissant à l'équation

$$\ddot{\alpha} + \Omega^2\,(z)\,\alpha = f\,(z, v), \tag{2.2}$$

où $u$ et $v$ sont des commandes. On admettra que la fréquence est assez élevée. Prenons par exemple

$$\Omega = \omega/\varepsilon. \tag{2.3}$$

Introduisons les variables de Van der Pol

$$\ddot{\alpha} = x \cos y, \qquad \dot{\alpha} = -\,\Omega x \sin y. \tag{2.4}$$

La condition de compatibilité des formules (2.4) s'écrit

$$\dot{x} \cos y - x\dot{y} \sin y = -\,\Omega\, x \sin y. \tag{2.5}$$

En dérivant la deuxième expression (2.4) et en portant dans l'équation (2.2), on obtient encore une équation par rapport à $x$ et $y$ :

$$-\,\dot{x}\Omega \sin y - \dot{y}\Omega x \cos y = \dot{\Omega} x \sin y + f\,(z, v) - \Omega^2 x \cos y. \tag{2.6}$$

23—01523

La résolution des équations (2.5) et (2.6) par rapport à $\dot{x}$ et $\dot{y}$ nous conduit au système d'équations

$$\begin{aligned} \dot{x} &= -\frac{f(z, v)}{\Omega}\sin y - \frac{\dot{\Omega}}{\Omega} x \sin^2 y, \\ \dot{y} &= \Omega - \frac{1}{\Omega}\left(\sin y \cos y \dot{\Omega} + \frac{f(z, v)}{x}\cos y\right). \end{aligned} \tag{2.7}$$

La relation (2.3) aidant et compte tenu de

$$\dot{\Omega} = \frac{d\Omega}{dz}\frac{dz}{dt} = \omega' Z, \quad \omega' = \frac{d\omega}{dz},$$

on ramène le système (2.7) à la forme

$$\dot{x} = \varepsilon X(x, y, z, u, v), \qquad \dot{y} = \Omega + \varepsilon Y(x, y, z, u, v), \tag{2.8}$$

où

$$\begin{aligned} X &= -\frac{f(z, v)}{\omega}\sin y - \frac{\omega'}{\omega} x \sin^2 y Z, \\ Y &= -\frac{1}{\omega}\left(\sin y \cos y \omega' Z + \frac{f(z, v)}{x}\cos y\right). \end{aligned}$$

L'introduction des variables de Van der Pol ramène l'équation (2.1) à la forme

$$\dot{z} = \varepsilon Z(z, x \cos y, u). \tag{2.9}$$

Si l'on admet que $u$ et $v$ sont des quantités données, alors le système (2.8), (2.9) se rapporte aux systèmes standards où les variables $z$ et $x$ sont lentes et $y$, rapide. A noter que les seconds membres du système d'équations (2.8), (2.9) sont des fonctions $2\pi$-périodiques de la variable rapide $y$.

On peut formuler divers problèmes variationnels pour le système (2.8), (2.9) et les plus fréquents d'entre eux contiendront en premier lieu des fonctionnelles de variables lentes. Attardons-nous sur ces problèmes. Sans nuire à la généralité on traitera les problèmes dont la fonctionnelle terminale est linéaire.

On étudiera donc des problèmes variationnels de la forme

$$J(u) = (c, x(T)) \Rightarrow \min \tag{2.10}$$

sous les conditions

$$\dot{x} = \varepsilon X(x, y, u), \qquad \dot{y} = \omega(x) + \varepsilon Y(x, y, u), \tag{2.11}$$

où $x$ (la variable lente) est une fonction vectorielle de dimension $n$, $y$, une fonction scalaire. Le vecteur $c$ est défini par ses coordonnées $c^{(i)}$ qui ne sont pas toutes nulles. Des conditions subsidiaires peuvent être imposées à l'extrémité droite de la trajectoire.

REMARQUE. Ce qui est important ici c'est que la commande peut intervenir aussi bien dans les équations « rapides » que dans les équations « lentes ». Ainsi, dans le cas du système avion-pilote automatique, c'est d'abord la variable rapide qui est commandée : un braquage des gouvernes d'altitude, par exemple, modifie l'angle d'attaque autour duquel oscille l'avion. La variable lente (la trajectoire) sera modifiée aussi, mais ce sera par voie de conséquence. Dans les systèmes économiques ou écologiques, c'est la situation inverse qui prévaut : on agit sur le fond lentement variable, par exemple en changeant la structure des investissements. ⊔.

b) *Moyennisation directe*. Soit donc à résoudre le problème (2.10), (2.11). La voie la plus légitime est de tenter de simplifier le problème primitif avant même de chercher l'extrémum de la fonctionnelle (2.10). Ce qui pose toute une série de problèmes spécifiques. En effet, supposons tout d'abord que la fonction $u$ est une fonction donnée du temps (ou des coordonnées de phase). Par exemple, $u = u(t)$. Le système (2.11) devient alors

$$\begin{aligned} \dot{x} &= \varepsilon X(x, y, u(t)) = \varepsilon \hat{X}(x, y, t), \\ \dot{y} &= \omega(x) + \varepsilon Y(x, y, u(t)) = \omega(x) + \varepsilon \hat{Y}(x, y, t). \end{aligned} \tag{2.12}$$

Les seconds membres du système (2.12) contiennent explicitement encore une variable rapide : le temps, et, de plus, les fonctions $\hat{X}$ et $\hat{Y}$ ne sont pas des fonctions périodiques de $t$ dans le cas général. On ne peut donc analyser ce système à l'aide de l'appareil exposé au chapitre IV pour la séparation des mouvements. Il est nécessaire de modifier sensiblement cette théorie et de l'adapter en conséquence. Il existe cependant une autre méthode qui consiste à assujettir les variations des commandes à des conditions subsidiaires. Supposons, par exemple, que la commande $u$ est une fonction du temps (et de la variable lente) variant lentement :

$$\dot{u} = \varepsilon U(x, t). \tag{2.13}$$

Alors le système (2.11), (2.12), (2.13) peut être traité par les méthodes du chapitre IV, même si formellement il n'est pas de la forme standard. Plus exactement, on cherchera les variables $x$ et $y$ sous la forme

$$x = \bar{x} + \varepsilon\xi(\bar{x}, \bar{y}, u), \qquad y = \bar{y} + \varepsilon\eta(\bar{x}, \bar{y}, u), \tag{2.14}$$

où $\bar{x}$ et $\bar{y}$ sont solutions des équations

$$\dot{\bar{x}} = \varepsilon A(\bar{x}, \bar{y}, u), \quad \dot{\bar{y}} = \omega(\bar{x}) + \varepsilon B(\bar{x}, \bar{y}, u). \tag{2.15}$$

En se servant de la transformation (2.14) et des formules (2.15) on peut développer une méthode d'approximations successives exactement comme au chapitre IV. Ce faisant, on remplace le système

primitif d'équations par un autre qui fera l'objet d'une moyennisation par rapport à la variable rapide $y$. Si $\omega$ dépend de $x$, alors le système de première approximation sera

$$\begin{aligned}\dot{\bar{x}} &= \varepsilon \overline{X}(\bar{x}, u),\\ \dot{\bar{y}} &= \omega(\bar{x}),\\ \dot{u} &= \varepsilon U(t, \bar{x}).\end{aligned} \tag{2.16}$$

Si $\omega$ ne dépend pas de $x$, alors pour déterminer toutes les variables avec la même précision, il faut remplacer la deuxième équation du système (2.16) par l'équation suivante:

$$\dot{\bar{y}} = \omega + \varepsilon \overline{Y}(\bar{x}, u). \tag{2.17}$$

La méthode qui permet de ramener le système primitif (2.11), (2.13) au système (2.16) s'appelle *méthode de moyennisation partielle*. Si l'on étudie le système (2.11), (2.13) dans lequel la fonction $U$ est une fonction du temps supposée donnée, alors la moyennisation partielle est asymptotique et sa justification ne pose pas de problème.

Ainsi, la méthode d'étude du problème de commande optimale pour le système (2.11) consiste à ramener ce système à un système tronqué de la forme (2.15) ou (2.16) (l'équation pour $\bar{y}$ peut être de la forme (2.17)). On résout ensuite le problème variationnel (2.10) sous les nouvelles conditions.

Dans quelle mesure cette opération est-elle justifiée? La réponse à cette question implique des études spéciales et complexes. Signalons tout d'abord que la proximité des solutions des systèmes (2.11), (2.13) et (2.16) n'entraîne pas forcément celle des valeurs optimales correspondantes de la fonctionnelle (2.10). En effet, l'affirmation que le système (2.16) est asymptotique se base sur l'hypothèse que la commande $u$ est non seulement une fonction connue, mais est d'une forme spéciale décrite par les formules (2.13), c'est-à-dire est une fonction du temps lentement variable. La fonction $u$ réalisant l'extrémum de la fonctionnelle dans le problème primitif de commande optimale peut être d'une tout autre nature. De plus, le système d'équations moyennisées en première approximation peut ne pas contenir telle ou telle commande. Donc, on ne peut parler du caractère strict de la moyennisation partielle pour la résolution des problèmes de commande optimale et il est aisé d'exhiber des exemples où la procédure décrite nous donne des commandes et une valeur de la fonctionnelle qui ne peuvent en aucune façon être regardées comme des approximations asymptotiques. Malheureusement, on ne connaît à ce jour aucun travail contenant des estimations de la fonctionnelle calculée par une procédure de moyennisation partielle (en fonction de $\varepsilon$).

Il existe néanmoins de nombreux problèmes qui ont été résolus avec succès par une moyennisation des équations initiales par rapport à la variable rapide $y$. La raison en est avant tout dans le fait que les extrémums des fonctionnelles sont généralement assez flous et il n'est pas nécessaire de résoudre ces problèmes avec une précision élevée. En vérité, il faut qu'une condition soit encore remplie: le système primitif doit être remplacé par un système approché qui contienne toutes les commandes. Si, par exemple, la variable commandée est la variable rapide $y$ et si le paramètre $\omega$ ne dépend pas de $x$, on peut alors se limiter à la première approximation, sinon, c'est-à-dire si $\omega = \omega(x)$, il faudra recourir à une deuxième approximation, puisqu'en première approximation le système risque de ne pas être commandé. Si la variable qui n'est pas commandée est la variable rapide, on peut toujours se servir de la seule première approximation.

Nous avons vu par ailleurs que la moyennisation n'a un sens que si les commandes sont des fonctions lentement variables. Donc, après avoir résolu le problème à l'aide du système (2.16), on doit s'assurer que les dérivées des fonctions $u(t)$ cherchées sont bien petites. Si tel est le cas, on peut espérer que la commande trouvée par une moyennisation du système (2.11) est une commande admissible, proche de l'optimale quant à la valeur de la fonctionnelle.

Il est possible que les dérivées des commandes soient grandes, ou même infiniment grandes dans la mesure où les commandes optimales seront continues par morceaux. Dans de tels cas il faut envisager les commandes approchées. Soit, par exemple, un scalaire $u$. Introduisons une nouvelle commande $\varphi(t)$ à l'aide de l'équation

$$\dot{u} = \varepsilon\varphi(t), \tag{2.18}$$

où $|\varphi| \leqslant \varphi^*$ et traitons $u(t)$ comme une coordonnée de phase. Supposons maintenant qu'au lieu du problème d'optimisation initial nous sachions résoudre un problème de commande optimale dans lequel le rôle des commandes cherchées est tenu par la fonction $\varphi(t)$ pour toute valeur $\varphi^*$.

On obtient en définitive les valeurs de la fonctionnelle $J$ en fonction de $\varphi^*$:

$$J = J(\varphi^*).$$

Si la fonction $J(\varphi^*)$ ne varie pas fortement et si la valeur $J(\varphi^*)$ est voisine de celle trouvée antérieurement à l'aide du système (2.16), alors la procédure de moyennisation utilisée est justifiée.

REMARQUE. La construction des commandes approchées par résolution de problèmes variationnels auxiliaires est manifestement assez laborieuse. Mais on peut utiliser n'importe quelle procédure d'approximation d'autant plus que la commande approchée a déjà été trouvée. La seule condition à poser est que

$$|\dot{u}| \leqslant \varepsilon\varphi^*. \tag{2.19}$$

Supposons donc qu'on soit passé au système moyennisé (2.16). Ecrivons les conditions nécessaires sous forme de principe du maximum. Cherchons le hamiltonien sous la forme (pour fixer les idées on considère le cas où $\omega$ ne dépend pas de $x$)

$$H = \varepsilon (\psi, \bar{X} (\bar{x}, u)) + \lambda\omega + \varepsilon\lambda\bar{Y} (x, u), \tag{2.20}$$

où $\psi$ est le vecteur des impulsions relatif à la variable $\bar{x}$, et $\lambda$, l'impulsion relative à la variable $\bar{y}$. Le hamiltonien ne contenant pas $\bar{y}$, on a $\dot{\lambda} = - \partial H/\partial \bar{y} = 0$, d'où $\lambda$ = const. Si par ailleurs la fonctionnelle terminale ne dépend pas de $y$ et $\lambda (T)$ n'est soumis à aucune condition, alors $\lambda (T) = 0$, et par suite $\lambda \equiv 0$. Ceci étant, le hamiltonien $H$ ne dépend pas de la commande figurant dans l'équation pour la variable rapide et est de la forme

$$H = \varepsilon (\psi, \bar{X} (\bar{x}, u)).$$

Donc, dans ce cas le problème se ramène uniquement à la commande d'un mouvement lent. Par conséquent, la méthode de moyennisation partielle ne peut être appliquée à la commande des mouvements rapides. Mais si l'on veut commander uniquement le mouvement lent à l'aide de la fonction $u$ et si la fonctionnelle est de la forme (2.10), c'est-à-dire $J = (c, x (T))$, alors le problème variationnel se ramène au problème aux limites suivant:

$$\begin{gathered} \dot{x} = \varepsilon\bar{X} (x, u), \\ \dot{\psi} = - \varepsilon\psi\bar{X}_x (x, u), \\ x (0) = x_0, \quad \psi (T) = - c, \end{gathered} \tag{2.21}$$

où la commande $u$ est définie à partir du principe du maximum

$$(\psi, \bar{X} (x, u)) \Rightarrow \max_{u \in G_u}. \tag{2.22}$$

Etant donné que le système d'équations (2.21) ne contient que les variables lentes, la résolution numérique de ce problème est assez stable et peut être effectuée à l'aide de la méthode de Krylov-Tchernoousko. Ceci étant, si le système d'équations primitif est d'ordre $n + 1$ ($n$ étant la dimension du vecteur des variables lentes), le système (2.21) sera d'ordre $2n$.

c) *Analyse d'un système pontriaguinien.* On peut aborder l'analyse du problème considéré sous un autre angle. Au lieu de simplifier le système et d'étudier ensuite le problème de commande, on peut d'abord écrire les conditions nécessaires sous la forme du principe du maximum, réduire le problème primitif à un problème aux limites et résoudre approximativement ce dernier par la méthode de moyennisation. Cette approche est notamment envisagée dans les travaux [27, 28].

Composons le hamiltonien pour le système (2.11):

$$H = \varepsilon\,(\psi,\ X\,(x,\ y,\ u)) + \lambda\omega + \varepsilon\lambda Y\,(x,\ y,\ u), \tag{2.23}$$

où $\psi$ et $\lambda$ ont la même signification que dans le numéro précédent. La commande $u$ peut être définie à partir du principe du maximum

$$H\,(x,\ y,\ \psi,\ \lambda,\ u) \Rightarrow \max_{u \in G_u}. \tag{2.24}$$

ce qui nous donne

$$u = U_1\,(x,\ y,\ \psi,\ \lambda,\ \varepsilon). \tag{2.25}$$

Les variables adjointes $\psi$ et $\lambda$ satisfont les équations

$$\begin{aligned} \dot{\psi} &= -\frac{\partial H}{\partial x} = -\{\varepsilon\psi X_x + \lambda\omega_x + \varepsilon\lambda Y_x\}, \\ \dot{\lambda} &= -\frac{\partial H}{\partial y} = -\varepsilon\{\psi X_y + \lambda Y_y\}. \end{aligned} \tag{2.26}$$

Vu que nous avons convenu d'étudier la fonctionnelle terminale sous la forme (2.10), les conditions de transversalité s'écrivent

$$\psi\,(T) = -c, \qquad \lambda\,(T) = 0. \tag{2.27}$$

Nous avons appelé le système

$$\dot{x} = \frac{\partial H}{\partial \psi}, \quad \dot{y} = \frac{\partial H}{\partial \lambda}, \quad \dot{\psi} = -\frac{\partial H}{\partial x}, \quad \dot{\lambda} = -\frac{\partial H}{\partial y}$$

système pontriaguinien (ou $\pi$-système). Les formules (2.26) montrent que le $\pi$-système n'est plus un système à phase tournante, puisque l'impulsion $\psi$ ne peut être traitée comme une variable lente: le second membre de l'équation pour $\psi$ contient la quantité $\lambda\omega_x$ qui peut être de l'ordre de l'unité. Le $\pi$-système ne peut non plus être regardé comme un système à deux phases tournantes, car les seconds membres du système (2.11), (2.26) ne seront pas des fonctions périodiques de l'impulsion $\psi$ lorsque la commande sera remplacée par son expression (2.25).

Néanmoins les conditions de transversalité (2.27) (c'est-à-dire le fait que $\lambda\,(T) = 0$) nous permettent de ramener le système (2.11), (2.26) à la forme standard des systèmes à phase tournante. En effet, étant donné que les systèmes envisagés sont autonomes (ne dépendant pas explicitement du temps), le hamiltonien est une intégrale première, c'est-à-dire que le long de la trajectoire du système (2.11) est réalisée la condition

$$H\,(x,\ y,\ \psi,\ \lambda,\ \varepsilon) = \text{const.}$$

Cette condition peut être mise sous la forme suivante:

$$H\,(x,\ y,\ \psi,\ \lambda,\ \varepsilon)\,|_t = H\,(x,\ y,\ \psi,\ \lambda,\ \varepsilon)\,|_T.$$

En vertu des conditions de transversalité (2.27) et de (2.23) on aura pour tout $u \in G_u$

$$H\mid_t = \varepsilon\,(\psi,\ X)\mid_{t=T}.$$

Cela signifie que dans les problèmes de commande optimale de systèmes à phase tournante les valeurs du hamiltonien sont de l'ordre de $\varepsilon$. Nous pouvons donc mettre l'expression de $H$ sous la forme

$$\varepsilon\,(\psi,\ X) + \lambda\omega + \varepsilon\lambda Y = \varepsilon h, \tag{2.28}$$

où $h$ est une constante.

La condition (2.28) permet d'exclure la variable $\lambda$:

$$\lambda = \frac{\varepsilon\,(h-(\psi,\ X))}{\omega+\varepsilon Y} = \varepsilon q\,(x,\ y,\ \psi,\ h,\ \varepsilon), \tag{2.29}$$

c'est-à-dire que $\lambda$ est de l'ordre de $\varepsilon$. L'expression (2.29) ne contient pas la commande, puisque nous avons admis qu'elle a été éliminée à l'aide de la formule (2.25). En excluant la commande des autres équations et en se servant de la formule (2.29), on ramène le système (2.11), (2.26) à la forme

$$\begin{aligned} \dot{x} &= \varepsilon X_1\,(x,\ y,\ \psi,\ h,\ \varepsilon), \\ \dot{y} &= \omega\,(x) + \varepsilon Y_1\,(x,\ y,\ \psi,\ h,\ \varepsilon), \\ \dot{\psi} &= -\varepsilon\,\frac{\partial}{\partial x}\,(\psi,\ X_1) - \varepsilon\omega_x q + O\,(\varepsilon^2). \end{aligned} \tag{2.30}$$

Le système (2.30) est un système classique à phase tournante. Il est d'ordre $2n + 1$ avec les conditions aux limites suivantes:

$$x\,(0) = x_0, \qquad y\,(0) = y_0, \qquad \psi\,(T) = -c. \tag{2.31}$$

Ainsi l'intégrale première $H = \varepsilon h$ nous a permis d'abaisser l'ordre du système d'une unité, mais la constante $h$ est inconnue. Donc, la solution du problème aux limites (2.30), (2.31) dépendra de $h$ comme d'un paramètre: $x = x\,(t,\ h)$, $y = (t,\ h)$, $\psi = \psi\,(t,\ h)$. Pour déterminer la constante $h$, on forme l'équation

$$H\mid_T = H\,(x\,(T,\ h), \quad y\,(T,\ h),\ \psi\,(T,\ h),\ \varepsilon) = h. \tag{2.32}$$

C'est une équation transcendante compliquée dans laquelle les fonctions $x$, $y$ et $\psi$ se définissent au moyen des solutions du problème aux limites (2.30), (2.31). Les conditions d'existence d'une solution sont, comme pour toute équation transcendante, probablement difficiles. Mais si l'on suppose que le problème primitif de commande optimale admet une solution, alors l'équation (2.32) devra avoir au moins une solution. Bien plus, dans le travail [31] on montre que l'équation (2.32) admet toujours plusieurs solutions. Ceci est dû au caractère oscillant de la fonction $y$.

Ainsi, le problème aux limites (2.30), (2.31) admet plusieurs solutions. Donc, pour déterminer la solution du problème de commande optimale, il faut choisir la racine de l'équation (2.32) à laquelle est associé le minimum de la fonctionnelle.

Le système (2.30) étant un système classique à phase tournante, on peut le moyenniser par rapport à $y$, c'est-à-dire le remplacer par le système

$$\begin{aligned}
&\dot{\bar{x}} = \varepsilon \bar{X}_1 (\bar{x}, \bar{\psi}, h, \varepsilon),\\
&\dot{\bar{y}} = \omega (\bar{x}), \qquad\qquad (2.33)\\
&\dot{\bar{\psi}} = -\varepsilon \bar{\psi} \bar{X}_{1x} (\bar{x}, \bar{\psi}, h, \varepsilon) - \varepsilon \omega_x (\bar{x}) \bar{q} (\bar{x}, \bar{\psi}, h, \varepsilon).
\end{aligned}$$

Les conditions aux limites pour le système (2.33) sont les conditions (2.31). L'équation pour $\bar{y}$ peut être intégrée indépendamment des autres. Donc, formellement on a un problème aux limites pour un système d'ordre $2n$. Mais le paramètre $h$ reste inconnu et pour le trouver il faut encore résoudre une équation transcendante (2.32). Cependant la situation se simplifie considérablement du fait que l'on doit résoudre cette équation à $\varepsilon$ près. Dans le travail [31] déjà cité, on montre que ces racines diffèrent entre elles d'une quantité de l'ordre de $\varepsilon$, autrement dit il nous suffit de trouver la valeur approchée de l'une d'elles seulement.

Si l'on désigne la solution exacte du problème primitif par $\hat{x}$, $\hat{y}$, $\hat{u}$, $\hat{\psi}$, alors la procédure de moyennisation nous donne une erreur définie par les formules

$$\begin{gathered}
\hat{x} = \bar{x} + O(\varepsilon), \quad \hat{y} = \bar{y} + O(\varepsilon), \quad \hat{\psi} = \bar{\psi} + O(\varepsilon),\\
\hat{u} = u + O(\varepsilon); \qquad\qquad (2.34)
\end{gathered}$$

de plus, l'approximation (2.34) est uniforme sur l'intervalle $[0, T]$ tout entier de variation de la variable indépendante $t$ et la quantité $T$ est d'autant plus grande que $\varepsilon$ est petit : $T = O(1/\varepsilon)$.

Illustrons l'approche exposée sur la résolution de deux problèmes simples empruntés aux travaux [27, 28] déjà cités. A noter que la procédure de moyennisation est considérablement simplifiée par la propriété d'une vaste classe de $\pi$-systèmes (admettant des solutions glissantes et singulières) d'être moyennisés comme suit : on moyennise d'abord le hamiltonien, puis on met l'équation sous la forme hamiltonienne en remplaçant $H$ par sa valeur moyennisée.

1. *Problème de l'excitation paramétrique optimale.* Un système oscillant est régi par l'équation

$$\ddot{\alpha} + (1 - \varepsilon u)\alpha = 0, \qquad (2.35)$$

où $\alpha$ est la coordonnée de phase. On demande de trouver sur un intervalle fixe $[0, T]$ une commande $u(t)$ vérifiant la condition $0 \leqslant u(t) \leqslant 1$ et réalisant à l'instant final $T$ le maximum de l'énergie complète du système

$$g(T) = \frac{1}{2}[\alpha^2(T) + \dot{\alpha}^2(T)].$$

Les variables de Van der Pol

$$\alpha = x \cos y, \quad \dot{\alpha} = -x \sin y,$$

nous permettent de ramener l'équation (2.35) au système classique

$$\begin{aligned} \dot{x} &= -\varepsilon u x \cos y \sin y, \\ \dot{y} &= 1 - \varepsilon u \cos^2 y, \end{aligned} \tag{2.36}$$

et de plus

$$g(T) = \frac{1}{2} x^2(T).$$

Donc, le problème primitif se ramène à la minimisation de la valeur finale de l'amplitude $x(T)$. Le hamiltonien du système (2.36) s'écrit

$$H = -\varepsilon\psi x u \cos y \sin y + \lambda - \varepsilon\lambda u \cos^2 y, \tag{2.37}$$

les variables adjointes satisfont le système

$$\begin{aligned} \dot{\psi} &= \varepsilon\psi u \cos y \sin y, \\ \dot{\lambda} &= \varepsilon x u \psi \cos 2y - 2\varepsilon\lambda u \cos y \sin y. \end{aligned} \tag{2.38}$$

Les conditions aux limites du $\pi$-système composé des équations (2.36) et (2.38) sont:

$$x(0) = x_0, \quad y(0) = y_0, \quad \psi(T) = -1, \quad \lambda(T) = 0. \tag{2.39}$$

La condition de maximum du hamiltonien $H$ défini par (2.37) nous conduit à l'équation

$$u = \theta(-\lambda - x\psi \operatorname{tg} y), \tag{2.40}$$

où $\theta(z) = 0$ pour $z < 0$ et $\theta(z) = 1$ pour $z > 0$ est la fonction de Heaviside. En portant (2.40) dans le hamiltonien (2.37) et en moyennisant sur $y$, on trouve

$$\overline{H} = \frac{1}{2\pi}\int_0^{2\pi} H\,dy = \lambda - \frac{\varepsilon}{2\pi}\left[x\psi + \frac{\lambda\pi}{2} + \lambda \operatorname{Arctg}\frac{\lambda}{x\psi}\right].$$

Mettons les équations sous la forme hamiltonienne. En tenant compte de la condition (2.39), on obtient en première approximation

$\lambda(t) \equiv 0$. Ceci simplifie la forme du système moyennisé:

$$\dot{x} = -\frac{\varepsilon}{2\pi} x, \quad \dot{y} = 1 - \frac{\varepsilon}{4}, \quad \dot{\psi} = \frac{\varepsilon}{2\pi} \psi.$$

En intégrant ce système en tenant compte de (2.39), on obtient

$$x(t) = x_0 \exp\left(-\frac{\varepsilon t}{2\pi}\right), \quad \psi(t) = -\frac{x(T)}{x(t)}, \quad y(t) = y_0 + \left(1 - \frac{\varepsilon}{4}\right) t.$$

La substitution de la solution trouvée dans (2.40) nous donne la loi de variation de la commande sous forme de synthèse:

$$u = \theta(\pm \operatorname{tg} y) = \theta(\mp y\dot{y}). \tag{2.41}$$

Dans les expressions (2.41) le signe supérieur correspond à une synthèse conduisant à une réduction optimale de l'amplitude, le signe inférieur, à un accroissement de l'amplitude.

2. *Optimisation des mouvements oscillatoires.* Les mouvements oscillatoires du pendule mathématique sont décrits par l'équation essentiellement non linéaire suivante:

$$\ddot{y} + u \sin y = 0. \tag{2.42}$$

On demande de trouver une commande $u(t)$, $u_1 \leqslant u \leqslant u_2$, telle que la vitesse angulaire soit minimale à la fin du mouvement. On admet que le système accomplit un mouvement oscillatoire durant le processus de commande. Suivant l'approche développée dans les chapitres précédents, introduisons les nouvelles variables $x$, $\tau$ et $\Omega$ à l'aide des relations

$$\frac{dy}{dt} = \Omega + x, \quad \tau = \frac{t}{\varepsilon}, \quad \Omega = \frac{1}{\varepsilon},$$

où $\Omega$ est la vitesse angulaire initiale qui est supposée grande, de sorte que $\varepsilon = \frac{1}{\Omega} \ll 1$ est un petit paramètre. Comme

$$\frac{d^2y}{dt^2} = \frac{d\dot{y}}{dt} = \frac{dx}{dt} = \frac{1}{\varepsilon} \frac{dx}{d\tau},$$

on est conduit au système d'équations suivant

$$\frac{dy}{d\tau} = 1 + \varepsilon x, \quad \frac{dx}{d\tau} = -\varepsilon u \sin y \tag{2.43}$$

qui est équivalent au système (2.42). Soit à minimiser la vitesse angulaire $\dot{y} = \Omega + x$ pour le système (2.43). En d'autres termes, la fonctionnelle $J$ est de la forme

$$J = x(T). \tag{2.44}$$

Le problème est donc de minimiser la fonctionnelle (2.44) sous les conditions (2.43). Désignons par $\psi$ et $\lambda$ les multiplicateurs de Lag-

range correspondant respectivement aux variables $x$ et $y$. Composons maintenant le hamiltonien:

$$H = -\psi\varepsilon \sin y + \lambda + \varepsilon\lambda x. \qquad (2.45)$$

Le principe du maximum $H \Rightarrow \max$ nous donne une condition pour la détermination de la commande:

$$u\psi \sin y \Rightarrow \min. \qquad (2.46)$$

De là on déduit immédiatement que

$$\text{si } \psi \sin y > 0, \quad \text{alors } u = u_1,$$
$$\text{si } \psi \sin y < 0, \quad \text{alors } u = u_2.$$

L'introduction de la fonction de Heaviside nous permet de regrouper ces deux conditions en une seule:

$$u = u_1\theta\,(\psi \sin y) + u_2\theta\,(-\psi \sin y).$$

Ecrivons maintenant les équations pour les variables adjointes:

$$\frac{d\psi}{d\tau} = -\frac{\partial H}{\partial x} = -\varepsilon\lambda; \quad \frac{d\lambda}{d\tau} = -\frac{\partial H}{\partial y} = \varepsilon\psi u \sin y. \qquad (2.47)$$

Le système (2.43), (2.47) est un système d'équations à une phase tournante dont les seconds membres sont des fonctions périodiques de la variable rapide $y$. Moyennisons la deuxième équation (2.47):

$$\frac{d\lambda}{d\tau} = \frac{\varepsilon\psi}{2\pi} \int_0^{2\pi} (u_1\theta\,(\psi \sin y) + u_2\theta\,(-\psi \sin y) \cos y\, dy.$$

Comme

$$\int_0^{\pi} \sin y\, dy = \int_{\pi}^{2\pi} \sin y\, dy = 0,$$

on a

$$\frac{d\lambda}{d\tau} = 0$$

et par suite

$$\lambda = \text{const}.$$

Or la condition de transversalité pour $\lambda$ est $\lambda\,(T) = 0$, donc

$$\lambda \equiv 0.$$

De là il s'ensuit que

$$\frac{d\psi}{d\tau} = -\varepsilon\lambda = 0.$$

La condition de transversalité $\psi(T) = -1$ nous donne finalement

$$\psi = -1,$$

d'où

$$u = u_1\theta(-\sin y) + u_2\theta(\sin y).$$

Moyennisons maintenant la deuxième équation (2.43):

$$\frac{dx}{d\tau} = -\frac{\varepsilon}{2\pi}\int_0^{2\pi} [u_1\theta(-\sin y) + u_2\theta(\sin y)] \sin y\, dy =$$

$$= -\frac{\varepsilon}{2\pi}\left\{\int_0^{\pi} u_2 \sin y\, dy + \int_{\pi}^{2\pi} u_1 \sin y\, dy\right\} = \frac{\varepsilon}{\pi}\{u_1 - u_2\}.$$

D'où il s'ensuit immédiatement

$$x = \frac{\varepsilon(u_1 - u_2)}{\pi}\tau.$$

Ceci achève la résolution du problème.

Cette méthode de résolution des problèmes d'optimisation est fréquemment utilisée actuellement. Cela s'explique par le fait que la résolution des problèmes d'optimisation dans les cas où les liaisons différentielles (les équations du mouvement) décrivent les variations de quantités oscillant rapidement pose des difficultés pratiquement insurmontables pour l'analyse numérique. Sans un traitement asymptotique il est peu probable que ces problèmes soient accessibles même si l'analyste dispose d'un ordinateur d'une puissance illimitée.

Dans ce paragraphe nous avons exposé assez brièvement quelques schémas possibles liés aux procédures de moyennisation dans les problèmes de commande optimale. Ces schémas peuvent être modi. fiés et perfectionnés grâce aux singularités des diverses équations-

## § 3. Problèmes singuliers de commande optimale

Le chapitre précédent était consacré à l'étude des méthodes d'analyse de systèmes non commandés contenant un petit paramètre en la dérivée:

$$\begin{aligned} \dot{x} &= X(x, y, t, \varepsilon), \\ \varepsilon\dot{y} &= Y(x, y, t, \varepsilon). \end{aligned} \tag{3.1}$$

Etant donné que les méthodes de calcul des trajectoires des systèmes de la forme (3.1) se basaient sur le théorème de Tikhonov, on a convenu d'appeler tikhonoviens les systèmes vérifiant les conditions du dit théorème. Dans ce paragraphe, on étudiera des systèmes iden-

tiques aux systèmes tikhonoviens mais contenant des fonctions libres (des commandes):

$$\dot{x} = X(x, y, t, u, v, \varepsilon), \quad \varepsilon\dot{y} = Y(x, y, t, u, v, \varepsilon), \qquad (3.2)$$

où $u$ et $v$ sont des commandes se trouvant à la disposition du Responsable associé au système. Cette forme d'écriture d'un système commandé contenant un paramètre est commode, car on peut y ramener une vaste classe de problèmes. Un cas particulier du système (3.2) nous est fourni par les systèmes de la forme

$$\dot{x} = X(x, y, t, u, \varepsilon), \quad \varepsilon\dot{y} = Y(x, y, t, v, \varepsilon). \qquad (3.3)$$

L'étude systématique des systèmes de la forme (3.2) n'en est qu'à ses balbutiements et les recherches mathématiques qui ont abouti à des résultats rigoureux ne fournissent pas en principe d'algorithmes effectifs; quant aux algorithmes effectifs composés pour résoudre les divers problèmes pratiques, ils ne sont pas rigoureusement justifiés. Pour rester fidèles aux objectifs de cet ouvrage, nous nous attarderons uniquement sur les problèmes relatifs au calcul effectif des commandes optimales pour les systèmes de la forme (3.2).

De même que dans les problèmes de commande optimale de systèmes à phase tournante, on peut proposer deux approches pour résoudre les problèmes d'optimisation pour les systèmes (3.2). La première consiste à simplifier le système initial et ensuite à résoudre le problème de commande optimale pour le système simplifié. La seconde, à composer directement le $\pi$-système en utilisant le fait qu'il contiendra de petits paramètres et ensuite à trouver des méthodes de résolution approchée des problèmes aux limites obtenus. Etudions successivement ces deux approches.

a) *Utilisation directe des procédures asymptotiques.* Rappelons le mécanisme général du calcul des trajectoires des systèmes tikhonoviens de la forme (3.1). En posant $\varepsilon = 0$, on compose le système générateur

$$\dot{x} = X(x, y, t, 0), \quad Y(x, y, t, 0) = 0, \qquad (3.4)$$

puis à partir de la deuxième équation du système (3.4), on exprime $y$ en fonction de $x$ et de $t$:

$$y(t) = y^0(x, t),$$

puis on porte $y^0(x, t)$ dans la première équation du système (3.4):

$$\dot{x} = X(x, y^0(x, t), t, 0). \qquad (3.5)$$

La solution du système (3.1) doit vérifier les conditions initiales:

$$x(0) = x_0, \quad y(0) = y_0. \qquad (3.6)$$

On peut satisfaire la première de ces conditions en résolvant le problème de Cauchy pour l'équation (3.5). Ceci nous donne la fonction $x^0(t)$ qui nous permet de composer la fonction

$$y^0 = y^0(x^0(t), t). \tag{3.7}$$

On pose

$$y = y^0 + z. \tag{3.8}$$

On linéarise la deuxième équation (3.1) par rapport à $z$ en y portant l'expression (3.8). En posant $\varepsilon = 1/\lambda$, on ramène cette équation à la forme

$$\dot{z} = \lambda A(t, \lambda) z + \lambda f(t, \lambda), \tag{3.9}$$

où la matrice $A$ et le vecteur $f$ sont des fonctions continues et dérivables du paramètre $1/\lambda$. En cherchant la solution de l'équation (3.9) qui vérifie la condition initiale

$$z(0) = y_0 - y^0(x_0, 0), \tag{3.10}$$

on satisfait la deuxième condition (3.6). On obtient l'intégrale générale de l'équation (3.9) et par suite la solution du problème de Cauchy (3.9), (3.10) sous forme de quadratures de fonctions variant lentement. En construisant la fonction frontière $z$, on peut préciser la fonction $x(t)$. Pour cela on pose

$$x = x^0 + x_1, \quad y = y^0 + z$$

dans la première équation du système (3.1) et on la linéarise par rapport à $x_1$ et $z$, la variable $x_1$ devant vérifier les conditions initiales nulles

$$x_1(0) = 0. \tag{3.11}$$

La solution asymptotique de ce problème peut également être acquise par des quadratures de fonctions lentement variables.

Ce schéma se généralise sans peine au système (3.2). Supposons pour fixer les idées que le problème de commande optimale consiste à minimiser la fonctionnelle terminale

$$J(u, v) = (c, x(T)) \Rightarrow \min. \tag{3.12}$$

Considérons au premier pas l'équation génératrice

$$Y(x, y, t, u, v, 0) = 0. \tag{3.13}$$

Sa solution $y$ sera une fonction dépendant non seulement de la variable de phase et du temps, mais aussi des commandes:

$$y = y^0(x, u, v, t). \tag{3.14}$$

En portant la fonction (3.14) dans la première équation du système (3.2) pour $\varepsilon = 0$, on résout le problème (3.12) sous la condition

$$x(0) = x_0.$$

REMARQUE. La construction de la fonction (3.14) peut être assez compliquée. Il est plus simple de considérer le problème

$$J(u, v) \Rightarrow \min, \qquad \dot{x} = X(x, y, t, u, v, 0) \tag{3.15}$$

sous la condition (3.13), en d'autres termes, de traiter la fonction $y$ comme une commande en lui imposant une condition subsidiaire du genre (3.13).

La résolution du problème (3.13), (3.15) nous donne les fonctions $x^0(t)$, $y^0(t)$, $u^0(t)$ et $v^0(t)$. La première question qui se pose est de savoir dans quelle mesure la prise en compte des fonctions frontières modifie la valeur de la fonctionnelle $J(u^0, v^0)$. Cette question est très importante sur le plan pratique. En effet, le système générateur est bien moins compliqué que le système primitif, son nombre de degrés de liberté peut être bien plus petit que celui du système primitif, la structure de ses seconds membres, bien plus simple : ils ne contiennent pas de fonctions variant rapidement. Donc, la résolution des problèmes d'optimisation à l'aide du système générateur peut être plus facile. Tous ces arguments plaident en faveur d'un passage au système générateur. Donc, le problème des estimations des commandes trouvées à l'aide des algorithmes simplifiés est un problème qui se pose essentiellement en pratique. Mais il est très épineux et il est rare qu'on lui trouve une réponse plus ou moins rigoureuse. Aussi, après avoir trouvé les commandes à l'aide des algorithmes rapides, peut-on envisager le problème relativement plus simple d'établir s'il est possible et nécessaire de préciser la solution $u^0(t)$ et $v^0(t)$ trouvée. A cet effet on pose $u = u^0(t)$, $v = v^0(t)$ et on reprend depuis le début la procédure de construction des fonctions frontières $x_1(t)$ et $z(t)$. Après avoir déterminé $x_1(t)$, on calcule l'écart

$$\Delta J = (c, x_1(T)). \tag{3.16}$$

Si la quantité $\Delta J$ n'est pas grande, il en est de même de la contribution, à la fonctionnelle, des fonctions frontières vérifiant la deuxième condition aux limites

$$y(0) = y_0, \tag{3.17}$$

et pour calculer les commandes il suffit d'utiliser les équations génératrices.

Signalons que $\Delta J$ est toujours $\geqslant 0$. En effet, tenir compte des fonctions frontières, c'est avant tout tenir compte des conditions supplémentaires (3.17), c'est-à-dire de la restriction de l'ensemble des stratégies admissibles.

Si $\Delta J$ est assez grande, la commande devrait être améliorée grâce aux fonctions frontières $z$ et $x_1$. A cet effet, il faut poser

$$u = u^0 + \delta u, \quad v = v^0 + \delta v,$$

porter ces expressions dans la deuxième équation (3.2) et linéariser cette dernière par rapport à $z$, $\delta u$ et $\delta v$. On obtient en définitive le

système linéaire

$$\dot{z} = \lambda A(t, \lambda) z + \lambda B(t, \lambda) \delta u + \lambda C(t, \lambda) \delta v + \lambda f(t, \lambda), \quad (3.18)$$

où $A$, $B$ et $C$ qui sont des matrices, $f$ qui est un vecteur sont des fonctions connues du temps. La variable $z$ doit satisfaire la condition initiale (3.10). La solution générale de l'équation (3.18) peut être acquise sous la forme explicite. Désignons par $Z(t)$ la matrice des solutions linéairement indépendantes de l'équation

$$\dot{z} = \lambda A^0(t) z,$$

où $A^0$ est le premier terme du développement de la matrice $A(t, \lambda)$:

$$A(t, \lambda) = A^0(t) + \lambda^{-1} A_1(t) + \ldots \quad (3.19)$$

En représentant les matrices $B$, $C$ et le vecteur $f$ sous la même forme (3.19) que $A$ et en se bornant aux premiers termes, on peut mettre l'intégrale générale du système (3.18) sous la forme

$$z = Zd - (A^0)^{-1} [B^0 \delta u + C^0 \delta v + f^0], \quad (3.20)$$

où $d$ est le vecteur des constantes arbitraires, qui se définit à partir de la condition (3.10). Mais nous ne pouvons encore pas l'expliciter, puisque les fonctions $\delta u$ et $\delta v$ sont *a priori* inconnues. En portant l'expression (3.20) dans la première équation (3.2), on trouve

$$\dot{x} = X(x, y^0 + Zd - (A^0)^{-1} [B^0 (u - u^0) + C^0 (v - v^0) + f^0], u, v, t, \varepsilon). \quad (3.21)$$

Pour cette équation, on doit résoudre le problème de commande optimale

$$J(u, v) = (c, x(T)) \Rightarrow \min \quad (3.22)$$

avec la condition subsidiaire (3.10), qui nous permettra de déterminer le vecteur $d$.

En étudiant les systèmes non commandés dans une situation analogue nous avons linéarisé l'équation en $x$ en posant $x = x^0 + x_1$. Ceci valait la peine, puisqu'on a pu représenter la solution du système linéaire obtenu, vérifiant la condition initiale $x_1(0) = 0$, par des formules finies. Ici nous avons affaire à un problème d'optimisation. La linéarisation nous conduit à un système linéaire assez volumineux. En appliquant le principe du maximum et en tenant compte des conditions

$$u = u^0 + \delta u \in G_u, \quad v = v^0 + \delta v \in G_v,$$

on aura à résoudre un problème aux limites qui sera aussi non linéaire. C'est pourquoi la linéarisation de l'équation (3.21) ne présente aucun intérêt. Il vaut mieux résoudre le problème (3.22) par une méthode numérique directe.

24—01523

b) *Assimilation d'une variable de phase à une commande.* Ce procédé a été proposé presque simultanément dans de nombreuses recherches consacrées au système

$$\begin{aligned} \dot{x} &= X(x, y, t), \\ \varepsilon\dot{y} &= Y(x, y, u, t). \end{aligned} \tag{3.23}$$

Les auteurs de ces recherches ne s'interrogeaient généralement pas sur le contenu mathématique des problèmes envisagés: ils avaient à résoudre d'intéressants problèmes techniques dont la position définissait implicitement, semble-t-il, la méthode de résolution. La voie dictée par l'intuition a conduit à des résultats qui répondaient parfaitement aux besoins posés.

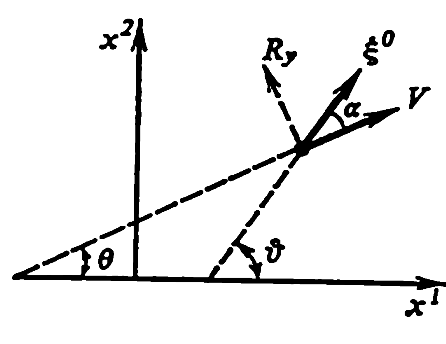


Fig. 6.1

Dans les années 50 le principal « bailleur de problèmes » de commande optimale était l'astronautique. Le rapide essor de la dynamique des missiles et véhicules spatiaux contribua au développement des méthodes d'optimisation. Montrons sur un exemple la teneur des problèmes posés. Cet exemple étant purement illustratif, on se bornera au cas d'un mouvement plan (fig. 6.1). On adopte les notations suivantes: $V$ est le vecteur vitesse du centre de masse du missile, $\xi^0$ le vecteur unitaire de son axe. L'angle $\alpha$ du vecteur vitesse et de l'axe du missile s'appelle angle d'attaque. Le mouvement du missile est déterminé par la pesanteur $mg$ qui est dirigée verticalement vers le bas, les forces aérodynamiques et la poussée propulsive.

Les forces aérodynamiques possèdent deux composantes: $R_x$ et $R_y$. La composante $R_x$ est de sens contraire au vecteur vitesse et est une fonction de $V^2$, $\alpha$ et $x^2$:

$$R_x = R_x(V^2, \alpha, x^2) = R_x((\dot{x}^1)^2 + (\dot{x}^2)^2, \alpha, x^2).$$

On l'appelle force de résistance frontale. La composante $R_y$ appelée portance est perpendiculaire au vecteur vitesse $V$. Elle est linéaire en $\alpha$:

$$R_y = K_y((\dot{x}^1)^2 + (\dot{x}^2)^2, x^2)\,\alpha.$$

La poussée propulsive $q$ est dirigée suivant le vecteur $\xi^0$ et s'exprime par

$$q = k_1 \frac{dm}{dt},$$

où $k_1$ est une constante (la vitesse d'écoulement des gaz). Les équations du mouvement du centre de masse du missile peuvent donc s'écrire sous la forme suivante (toutes les expressions des seconds membres du système (3.24) sont des quantités scalaires):

$$m\frac{d^2x^1}{dt^2}=R_x\cos\theta-R_y\sin\theta+q\cos\vartheta,$$
$$m\frac{d^2x^2}{dt^2}=R_x\sin\theta+R_y\cos\theta+q\cos\vartheta, \qquad (3.24)$$
$$\frac{dm}{dt}=\frac{q}{k_1},$$

où

$$\theta=\operatorname{Arc\,tg}\frac{\dot{x}^2}{\dot{x}^1}, \qquad \vartheta=\theta+\alpha.$$

En posant $\dot{x}^1=x^3$, $\dot{x}^2=x^4$, on peut mettre le système (3.24) sous la forme suivante dans laquelle la signification des notations $X^3$ et $X^4$ est parfaitement claire:

$$\dot{x}^1=x^3,\quad \dot{x}^2=x^4,$$
$$\dot{x}^3=X^3(x^2,\ x^3,\ x^4,\ \alpha), \qquad (3.25)$$
$$\dot{x}^4=X^4(x^2,\ x^3,\ x^4,\ \alpha).$$

On admettra dans cet exemple que la perte de masse $q$ est une quantité donnée. Ainsi, dans les missiles équipées d'un moteur à poudre, la perte de masse est définie par la loi de combustion de la poudre. Dans ce cas, la dernière équation du système (3.24) ne figure pas dans (3.25), puisqu'elle s'intègre indépendamment des autres, et l'on peut alors traiter la masse $m$ de la fusée comme une fonction connue du temps: $m=m(t)$. Le système (3.25) n'est pas fermé: il renferme encore une variable, l'angle d'attaque $\alpha$. Pour décrire les variations de cet angle, il faut composer l'équation du mouvement du missile par rapport à son centre de masse. Cette équation s'écrit

$$J\frac{d^2\vartheta}{dt^2}=M, \qquad (3.26)$$

où $J$ est le moment d'inertie, $M$, le moment des forces extérieures. Nous admettrons que le pilotage se fait par les gouvernes aérodynamiques (ou les intercepteurs). Le moment $M$ est alors la somme des moments des forces aérodynamiques $M_1$ et du moment $M_g$ des gouvernes. La poussée propulsive ne crée pas de moment, car elle est orientée le long de l'axe de symétrie du missile, lequel axe passe par le centre de masse. Pour le moment aérodynamique $M_1$, on a

$$M_1=-k(x^2,\ x^3,\ x^4)\,\alpha,\quad k>0.$$

Le moment $M_g$ est proportionnel à l'angle de rotation $u$ des gouvernes:

$$M_g = k_g (x^2, x^3, x^4)\, u.$$

On peut donc mettre l'équation (3.26) sous la forme

$$J \frac{d^2\vartheta}{dt^2} = J \left( \frac{d^2\alpha}{dt^2} + \frac{d^2\theta}{dt^2} \right) = -k\alpha + k_g u, \tag{3.27}$$

où

$$\frac{d^2\theta}{dt^2} = \frac{d^2}{dt^2} \operatorname{Arc\,tg} \frac{x^4}{x^3}.$$

En dérivant et en remplaçant les dérivées par leurs expressions (3.25), on peut mettre $d^2\theta/dt^2$ sous la forme d'une fonction des variables de phase $x^i$ et de l'angle d'attaque $\alpha$:

$$\frac{d^2\theta}{dt^2} = f(x^1, x^2, x^3, x^4, \alpha).$$

L'équation (3.27) devient maintenant

$$J \frac{d^2\alpha}{dt^2} = -k\alpha + k_g u - Jf. \tag{3.28}$$

Nous avons ainsi montré que le mouvement plan du missile peut être décrit par un système de la forme (3.23). En effet, le moment d'inertie $J$ est petit. Cela signifie que les oscillations propres d'équation

$$J \frac{d^2\alpha}{dt^2} + k\alpha = 0,$$

ont une période qui est petite en regard de la durée du vol du missile. Posons

$$J = \varepsilon^2 a, \qquad \varepsilon\dot{\alpha} = \beta$$

et mettons l'équation (3.28) sous forme du système

$$\varepsilon\dot{\alpha} = \beta, \quad \varepsilon\dot{\beta} = -\frac{k}{a}\alpha + \frac{k_g}{a} u - \varepsilon^2 f. \tag{3.29}$$

Nous sommes ainsi conduits au système (3.25), (3.29). La seule commande de ce système est l'angle de rotation $u$ des gouvernes. La commande n'agit que sur les variables du deuxième groupe, variables dont les équations contiennent de petits paramètres en les dérivées supérieures.

Revenons maintenant aux équations (3.23). Les innombrables travaux consacrés dans les années 50 et 60 à la mécanique du vol des missiles ont étudié des problèmes de commande optimale dans lesquels le missile était assimilé à un point matériel. Cette approche

a conduit d'étudier seulement la première équation

$$\dot{x} = X(x, y, t) \tag{3.30}$$

du système (3.23). Mais comme cette équation ne contient pas de commande, on a pris pour telle la variable de phase $y$, qui représente l'angle d'attaque dans notre exemple. La signification physique d'une telle schématisation du problème est évidente: puisque le moment d'inertie du missile est très petit et les gouvernes, assez efficaces, on peut négliger le mouvement relatif en estimant que l'angle d'attaque se définit (ou est donné) d'une manière ou d'une autre comme une fonction du temps et cette fonction peut être réalisée. La dernière proposition exprime justement la petitesse du moment d'inertie ou, plus exactement, la petitesse du rapport du moment d'inertie du missile au moment engendré par les gouvernes. Les gouvernes agissent pratiquement sans retard sur l'angle d'attaque en raison de la petitesse de ce rapport.

Dans le cadre de l'étude du mouvement des missiles, le problème de commande optimale qui se posait le plus naturellement était le problème de rapidité. Mais ce n'était pas le seul considéré. La résolution du problème de commande optimale pour le système (3.30) donnait la variable de phase $y$:

$$y = y^*(t). \tag{3.31}$$

Les auteurs de ces recherches s'arrêtaient généralement là. En réalité il fallait encore analyser la possibilité de réalisation de la loi de commande (3.31). A cet effet, il fallait considérer la deuxième équation du système (3.23) que, compte tenu de (3.31), nous écrirons sous la forme

$$\varepsilon \frac{dy^*}{dt} = Y(x^*, y^*, u, t), \tag{3.32}$$

où $x^*(t)$ désigne la trajectoire optimale. L'équation (3.32) sert à déterminer $u(t)$. La solution (si elle existe) de cette équation est précisément la commande optimale. En effet, toute cette procédure n'a servi qu'à nous dédouaner d'une contrainte: la deuxième équation du système (3.23), c'est-à-dire à élargir l'ensemble des stratégies admissibles et ensuite à montrer que la condition (3.23) est remplie. A noter que la commande doit satisfaire encore une contrainte de la forme

$$u(t) \in G_u, \tag{3.33}$$

et qu'il est possible qu'aucun élément de $G_u$ ne soit solution de l'équation (3.32). Dans ce cas il faut approcher la fonction $y^*(t)$ à l'aide de la solution de l'équation

$$\varepsilon \frac{dy}{dt} = Y(x, y, u, t). \tag{3.34}$$

Voici une éventuelle procédure de calculs. On pose

$$x = x^*, \qquad y = y^* + z,$$

puis on linéarise l'équation (3.34) sur $z$. On obtient ainsi l'équation

$$\varepsilon \frac{dz}{dt} = A(t, u) z + f(t, u) \tag{3.35}$$

pour laquelle on pose le problème

$$J(u) = z^2(T) \Rightarrow \min_{u \in G_u}. \tag{3.36}$$

La structure de la fonctionnelle (3.36) (qui mesure la qualité de l'approximation) est un problème assez délicat. En fait, la seule caractéristique devait être la fonctionnelle du problème primitif, mais la résolution directe de ce dernier est assez compliquée et réclame la mise au point de méthodes itératives. La résolution du problème variationnel auxiliaire (3.36) constitue précisément l'une des étapes de cette procédure itérative.

REMARQUE. Soit $\hat{z}$ la solution du problème (3.36). Si dans la fonctionnelle $I(y)$ du problème primitif, on remplace $y^*$ par $y^* + \hat{z}$, il est alors évident que

$$I(y^* + \hat{z}) \geqslant I(y^*),$$

puisque $y^*$ est la solution optimale du problème simplifié (3.30), autrement dit $I(y^*)$ est un minorant de la fonctionnelle du problème primitif. Le calcul de

$$\Delta I = I(y^* + \hat{z}) - I(y^*)$$

nous donne une estimation de l'écart par rapport à la valeur de la fonctionnelle primitive. L'adéquation des itérations ultérieures est laissée à l'appréciation de l'analyste à partir de l'estimation de $\Delta I$.

Ainsi le problème du choix de la fonctionnelle (3.36) est une pure affaire de commodité: la fonctionnelle (3.36) doit se prêter au calcul numérique du problème variationnel. On peut par exemple prendre la fonctionnelle quadratique

$$J(u) = \int_0^T (z, z)\, dz. \tag{3.37}$$

Signalons que la condition (3.33), *a fortiori* si l'ensemble $G_u$ est fermé, complique la procédure numérique; on aura donc intérêt à se dédouaner d'une manière ou d'une autre de cette contrainte, par exemple on remplacera la fonctionnelle (3.37) par la fonctionnelle

$$J(u) = \int_0^T [(z, z) + \lambda (u, Ru)]\, dt, \tag{3.38}$$

où $R$ est une matrice définie positive, par exemple la matrice unité. Le problème variationnel avec la fonctionnelle (3.38) et les contraintes (3.35) présente de nombreuses particularités sur lesquelles nous reviendrons.

Récapitulons. La méthode de résolution approchée du problème de commande optimale pour le système (3.23) comporte deux étapes.

1. On élimine d'abord l'équation en $\dot{y}$; la variable de phase $y$ qui devient une nouvelle commande est choisie de manière à minimiser la fonctionnelle du problème.

2. On résout ensuite un problème variationnel subsidiaire pour obtenir la commande qui approche le mieux (dans tel ou tel sens) la valeur optimale de $y$.

Une fois en possession d'une telle solution, on peut sans peine imaginer divers procédés pour améliorer la commande trouvée.

c) *Analyse d'un système pontriaguinien.* Dans les numéros précédents nous avons étudié des méthodes de résolution de problèmes de commande optimale pour des systèmes décrits par des équations de la forme (3.2). Ces méthodes étaient toutes fondées sur une simplification préalable des équations. Mais on sait que ce problème peut être abordé sous un angle différent : on peut écrire d'abord le $\pi$-système, ramener le problème de commande optimale à un problème aux limites et résoudre ensuite ce dernier en utilisant le fait que le $\pi$-système contient des paramètres.

Considérons le système

$$\dot{x} = X(x, y, u, t, \varepsilon), \quad \varepsilon\dot{y} = Y(x, y, u, t, \varepsilon) \tag{3.39}$$

et choisissons la commande de manière que

$$J = (c, x(T)) \Rightarrow \min. \tag{3.40}$$

Prenons les conditions initiales de la forme (3.6) et les contraintes sur la commande, de la forme (3.33).

Composons le hamiltonien

$$H = (\psi, X) + \frac{1}{\varepsilon}(\varphi, Y), \tag{3.41}$$

où les impulsions $\psi$ et $\varphi$ sont solutions des équations

$$\dot{\psi} = -\frac{\partial H}{\partial x} = -X_x^*\psi - \frac{1}{\varepsilon} Y_x^*\varphi,$$
$$\dot{\varphi} = -\frac{\partial H}{\partial y} = -X_y^*\psi - \frac{1}{\varepsilon} Y_y^*\varphi,$$

que nous mettrons sous la forme suivante :

$$\begin{aligned} \varepsilon\dot{\psi} &= \Psi(x, y, \varphi, \psi, t, \varepsilon), \\ \varepsilon\dot{\varphi} &= \Phi(x, y, \varphi, \psi, t, \varepsilon). \end{aligned} \tag{3.42}$$

Les fonctions $\varphi$ et $\psi$ doivent satisfaire les conditions de transversalité

$$\psi(T) = -c, \quad \varphi(T) = 0. \tag{3.43}$$

Nous constatons que les équations pour les variables adjointes contiennent aussi des petits paramètres en les dérivées. Eliminons la commande $u$ à l'aide du principe du maximum:

$$\varepsilon(\psi, X) + (\varphi, Y) \Rightarrow \max_{u \in G_u}, \tag{3.44}$$

d'où

$$u = U(x, y, \varphi, \psi, t, \varepsilon).$$

En portant l'expression de $u$ dans les systèmes d'équations (3.39) et (3.42), on obtient le système d'équations

$$\begin{aligned} \dot{x} &= \hat{X}(x, y, \varphi, \psi, t, \varepsilon), \\ \varepsilon\dot{y} &= \hat{Y}(x, y, \varphi, \psi, t, \varepsilon), \\ \varepsilon\dot{\varphi} &= \hat{\Phi}(x, y, \varphi, \psi, t, \varepsilon), \\ \varepsilon\dot{\psi} &= \hat{\Psi}(x, y, \varphi, \psi, t, \varepsilon). \end{aligned} \tag{3.45}$$

Le problème de commande optimale est ainsi ramené à la recherche de fonctions $x$, $y$, $\varphi$ et $\psi$ satisfaisant le système (3.45) et les conditions aux limites

$$\begin{aligned} &t = 0, \quad x = x_0, \quad y = y_0, \\ &t = T, \quad \varphi = 0, \quad \psi = -c. \end{aligned} \tag{3.46}$$

La résolution numérique du problème aux limites (3.45), (3.46) peut poser quelques problèmes, puisque toutes les variables, hormis $x$, sont rapides. Discutons les éventuels procédés de résolution numérique de ce problème.

Suivant les raisonnements antérieurs, composons le système générateur. Posons

$$\begin{aligned} &\hat{Y}(x, y, \varphi, \psi, t, 0) = 0, \\ &\hat{\Phi}(x, y, \varphi, \psi, t, 0) = 0, \\ &\hat{\Psi}(x, y, \varphi, \psi, t, 0) = 0. \end{aligned} \tag{3.47}$$

La résolution du système (3.47) par rapport aux variables rapides $y$, $\varphi$ et $\psi$ nous donne

$$y = y^0(x, t), \quad \varphi = \varphi^0(x, t), \quad \psi = \psi^0(x, t).$$

Portons ces fonctions dans la première équation du système (3.45) dans lequel nous poserons $\varepsilon = 0$:

$$\dot{x} = \hat{X}(x, y^0(x, t), \varphi^0(x, t), \psi^0(x, t), t, 0) = \hat{X}^0(x, t). \tag{3.48}$$

Cherchons la solution de l'équation (3.48) qui satisfait la condition $x(0) = x_0$, c'est-à-dire la première condition initiale (3.46). Désignons cette solution par $x^0(t)$. En vertu de la théorie générale exposée au chapitre V, la fonction $x^0(t)$ approche uniformément la solution exacte $x(t)$ sur l'intervalle $[0, T]$ tout entier:

$$x(t) = x^0(t) + O(\varepsilon).$$

La réalisation numérique des procédures de formation de la fonction $x^0(t)$ rencontre des obstacles de taille sur lesquels nous allons justement nous arrêter. Posons $t = 0$, $\varepsilon = 0$ et écrivons le système d'équations (3.39), (3.42) et la condition (3.44):

$$\dot{x}^0 = X(x^0, y, u, t, 0),$$
$$Y(x_0, y_0, u_0, t_0, 0) = 0, \qquad (3.39')$$
$$\Psi(x_0, y_0, \psi_0, \varphi_0, u_0, t_0, 0) = 0;$$
$$\Phi(x_0, y_0, \psi_0, \varphi_0, u_0, t_0, 0) = 0, \qquad (3.42')$$
$$H(x_0, y_0, \varphi_0, \psi_0) \Rightarrow \max_{u_0 \in U}.$$

Dans ce système d'équations nous ne connaissons que l'état initial $x_0$. Donc, quel que soit le schéma aux différences retenu, le premier pas de la procédure numérique consistera à chercher le maximum de $H$ sous les conditions (3.39′) et (3.42′). Nous conviendrons d'appeler ce problème *problème de compatibilité des conditions initiales.* Les difficultés s'estomperont aux étapes suivantes de la résolution du problème de Cauchy. Montrons-le sur l'exemple du schéma élémentaire d'Euler. Les calculs peuvent être conduits d'après le schéma suivant:

$$x_1 = x_0 + \tau X(x_0^0, y_0, u_0, t_0, 0),$$
$$y_1 = Y(x_1, y_0, u_0, t_0, 0) - y_0,$$
$$\psi_1 = \Psi(x_1, y_0, \psi_0, \varphi_0, t_0, 0) - \psi_0,$$
$$\varphi_1 = \Phi(x_1, y_0, \psi_0, \varphi_0, t_0, 0) - \varphi_0.$$

Il reste bien sûr le problème de déterminer la commande, mais celui-ci ne contiendra plus de conditions de type liaisons, puisque les quantités $x_1$, $y_1$, $\psi_1$ et $\varphi_1$ sont déjà connues.

Pour approcher les autres variables uniformément et avec la même précision, il nous faut construire les fonctions frontières. Reprenons quelques raisonnements du chapitre précédent. Posons

$$y = y^0 + y_1, \quad \varphi = \varphi^0 + \varphi_1, \quad \psi = \psi^0 + \psi_1.$$

Portons ces expressions dans les trois dernières équations du système (3.45) et linéarisons les seconds membres par rapport à $y_1$, $\varphi_1$ et

$\psi_1$. Nous obtenons en définitive ($\lambda = 1/\varepsilon$):

$$\begin{aligned}\dot{y}_1 &= \lambda A_{11}(t, \lambda)\, y_1 + \lambda A_{12}(t, \lambda)\, \varphi_1 + \\ &\qquad + \lambda A_{13}(t, \lambda)\, \psi_1 + \lambda f_1(t, \lambda),\\ \dot{\varphi}_1 &= \lambda A_{21}(t, \lambda)\, y_1 + \lambda A_{22}(t, \lambda)\, \varphi_1 + \\ &\qquad + \lambda A_{23}(t, \lambda)\, \psi_1 + \lambda f_2(t, \lambda),\\ \dot{\psi}_1 &= \lambda A_{31}(t, \lambda)\, y_1 + \lambda A_{32}(t, \lambda)\, \varphi_1 + \\ &\qquad + \lambda A_{33}(t, \lambda)\, \psi_1 + \lambda f_3(t, \lambda).\end{aligned} \tag{3.49}$$

Les conditions aux limites pour ce système d'équations différentielles linéaires sont

$$\begin{aligned}y_1(0) &= y_0 - y^0(x^0(0), 0),\\ \varphi_1(T) &= -\varphi^0(T),\\ \psi_1(T) &= -c - \psi^0(T).\end{aligned} \tag{3.50}$$

Le problème aux limites (3.49), (3.50) peut être intégré par la méthode de factorisation. Mais cette méthode passe difficilement, car les dérivées des fonctions du système (3.50) sont élevées. Il faut donc essayer de tirer parti de certaines particularités du système (3.39) à condition qu'il soit tikhonovien. Pour appliquer l'appareil du chapitre précédent, il faut que la solution de la deuxième équation du système (3.39) soit asymptotiquement stable pour des valeurs fixes de toutes les variables, hormis $y$. De là il s'ensuit que les parties réelles des valeurs propres de la matrice $A_{11}$ doivent être strictement négatives. Donc, les parties réelles des valeurs propres des matrices $A_{22}$ et $A_{33}$ seront strictement positives. En d'autres termes, lorsque $t$ croîtra, $y$ décroîtra de gauche à droite et $\varphi$ et $\psi$, de droite à gauche. Donc, les calculs ultérieurs devront être conduits de la manière suivante.

1. Suivant la théorie exposée au chapitre V, on construit explicitement l'intégrale générale du système (3.49). Cette intégrale contiendra $n + 2m$ constantes arbitraires ($n$ et $m$ étant les dimensions respectives des vecteurs $x$ et $y$).

2. On porte l'intégrale générale dans les conditions (3.50). On obtient $n + 2m$ équations pour la détermination des constantes arbitraires.

3. Vu qu'une partie des conditions sera satisfaite à gauche pour $t = 0$ et une partie, à droite pour $t = T$, le système d'équations linéaires par rapport aux constantes arbitraires contiendra des termes de grandeur très différente, et cette différence est d'autant plus élevée que $\lambda$ est grand. Ainsi, les termes contenant les valeurs propres de la matrice $A_{11}$ seront les « zéros de l'ordinateur » pour $t = T$ et on peut les négliger. De façon analogue, les termes définissant les fonctions frontières $\varphi_1$ et $\psi_1$ et dépendant des valeurs propres des matri-

ces $A_{22}$ et $A_{33}$ seront très petits pour $t = 0$. Ceci nous permet d'élaborer des procédures rationnelles de calcul numérique des constantes arbitraires.

Donc, l'application des méthodes du petit paramètre et de la technique d'analyse des systèmes tikhonoviens permet de simplifier considérablement les méthodes numériques de résolution des problèmes de commande optimale.

d) *Correction optimale dans les systèmes linéaires contenant un grand paramètre.* Au n° b) ci-dessus on a montré comment apparaît le problème de commande optimale pour les équations de la forme

$$\varepsilon \dot{y} = Y(y, t, u, \varepsilon). \quad (3.51)$$

On se penche ici sur le cas particulier du système (3.51) où le second membre est linéaire aussi bien en la commande $u$ qu'en la variable de phase $y$:

$$\dot{y} = \lambda A(t, \lambda) y + \lambda B(t, \lambda) u + \lambda f(t, \lambda), \quad (3.52)$$

où $\lambda = 1/\varepsilon$, $A$, $B$ et $f$ sont bornées pour $\lambda \to \infty$. Soit à minimiser une fonctionnelle de la forme (3.38) pour l'équation (3.52). Pour ne pas alourdir les calculs, on admettra que

$$I = \int_0^T [(y, y) + \mu(u, u)]\, dt. \quad (3.53)$$

Le problème de minimiser la fonctionnelle (3.53) sous les conditions (3.52) s'est déjà posé à nous en tant qu'étape de réalisation de la procédure de recherche d'une commande optimale dans des problèmes de forme plus générale. Mais ces problèmes présentent un intérêt en soi. Les problèmes de correction jouent un rôle important dans les procédures de commande. Ils se ramènent à l'analyse de systèmes d'équations linéaires de la forme

$$dy/dt = Ay + Bu + f, \quad (3.54)$$

où les matrices $A$ et $B$ et le vecteur $f$ sont des fonctions variant lentement:

$$A = A(\varepsilon, t), \quad B = B(\varepsilon, t), \quad f = f(\varepsilon, t).$$

L'emploi de la fonctionnelle quadratique (3.53) tient à plusieurs raisons. C'est d'abord la forme de fonctionnelle la plus simple qui soit justiciable de méthodes analytiques ou qui ramène le problème à un problème aux limites intégrable numériquement par des procédures régulières (du genre méthode de factorisation). On retrouve la fonctionnelle quadratique dans les problèmes traitant des techniques spatiales. Si la correction est effectuée avec un moteur nucléaire parfait, alors les dépenses d'énergie sont décrites, comme le montrent

de nombreux travaux, par une fonctionnelle quadratique (cf. [34]). Ainsi, dans ces problèmes, la minimisation d'une fonctionnelle de la forme

$$J = \int_0^T (u, u)\, dt$$

exprime qu'a été retenu le mode de correction le plus économique.

Revenons maintenant à la minimisation de la fonctionnelle (3.53) sous les conditions (3.54). En faisant le changement $\tau = \varepsilon t$ et en posant $\lambda = 1/\varepsilon$ dans le système (3.54), on le ramène à la forme (3.52). Composons le hamiltonien pour le problème (3.52), (3.53):

$$H = (\psi, \lambda Ay + \lambda Bu + \lambda f) - (y, y) - \mu (u, u); \qquad (3.55)$$

l'impulsion $\psi$ est solution du système d'équations

$$\dot{\psi} = -\lambda A^* \psi + 2y. \qquad (3.56)$$

Supposons que la commande $u$ n'est soumise à aucune contrainte; cette commande se définit alors à partir de la condition $\partial H/\partial u = 0$, où

$$\partial H/\partial u = \lambda B^* \psi - 2\mu u,$$

d'où

$$u = \frac{\lambda B^* \psi}{2\mu}. \qquad (3.57)$$

En portant l'expression (3.57) dans l'équation (3.52), on obtient

$$\dot{y} = \lambda Ay + \frac{\lambda^2}{2\mu} BB^* \psi + \lambda f. \qquad (3.52')$$

On voit sur la formule (3.57) que le paramètre $\mu$ dont nous sommes maîtres doit être choisi de telle sorte que la commande $u$ soit finie quel que soit $\lambda$. En d'autres termes, les quantités $\lambda$ et $\mu$ doivent être du même ordre de grandeur. Le plus simple est de faire $\lambda = \mu$, l'équation (3.52') devient alors

$$\dot{y} = \lambda Ay + \frac{1}{2} \lambda BB^* \psi + \lambda f. \qquad (3.58)$$

Nous sommes ainsi conduits à un problème aux limites pour le système (3.56), (3.58). Etudions quelques traits de ce problème sur un exemple élémentaire.

e) *Exemple illustratif.* Traitons le cas particulier du système (3.56), (3.58), où $y$ est un scalaire et mettons-le sous la forme

$$\dot{y} = -\lambda \nu y + \frac{1}{2} \lambda b^2 \psi + \lambda f, \quad \dot{\psi} = \lambda \nu \psi + 2y. \qquad (3.59)$$

REMARQUE. La quantité $\nu$ doit être strictement positive dans les équations (3.59) si l'on veut que les conditions du théorème de Tikhonov soient réalisées pour l'analogue non commandé du système étudié.

Construisons l'intégrale générale du système (3.59) en utilisant la technique d'intégration asymptotique. Considérons d'abord le système homogène

$$\dot{y} = -\lambda\nu y + \frac{1}{2}\lambda b^2\psi, \quad \dot{\psi} = \lambda\nu\psi + 2y \tag{3.60}$$

et cherchons sa solution sous la forme

$$\begin{aligned} y &= \exp\left\{\lambda \int_0^t \omega\, dt\right\} \left(y_1 + \frac{1}{\lambda} y_2 + \ldots\right), \\ \psi &= \exp\left\{\lambda \int_0^t \omega\, dt\right\} \left(\psi_1 + \frac{1}{\lambda} \psi_2 + \ldots\right). \end{aligned} \tag{3.61}$$

En portant les expressions (3.61) dans les équations (3.60) et en regroupant les puissances semblables de $\lambda$, on obtient le système d'équations algébriques suivant pour la détermination des fonctions inconnues $y_i$ et $\psi_i$:

$$(\omega + \nu) y_1 - \frac{1}{2} b^2\psi_1 = 0, \quad (\omega - \nu)\psi_1 = 0, \tag{3.62}$$

$$(\omega + \nu) y_2 - \frac{1}{2} b^2\psi_2 = -\dot{y}_1, \quad (\omega - \nu)\psi_2 = -\dot{\psi}_1 + 2y_1, \tag{3.63}$$

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

Considérons le système (3.62). C'est un système homogène de deux équations linéaires. Pour qu'il admette une solution il est nécessaire et suffisant que la fonction $\omega$ soit racine de l'équation caractéristique. Cette équation admet deux racines

$$\omega_1 = -\nu, \quad \omega_2 = +\nu.$$

Traitons d'abord le cas où $\omega = \omega_1 = -\nu$. On déduit alors de la deuxième équation (3.62) que

$$2\nu\psi_1 = 0.$$

Comme $\nu \neq 0$, il vient

$$\psi_1 = \psi_{11} = 0. \tag{3.64}$$

Dans ce cas $y_1 = y_{11}$ est une quantité arbitraire. Pour la déterminer il nous faut considérer les équations (3.63), qui, en vertu de (3.64), s'écrivent

$$-\frac{1}{2} b^2\psi_{21} = -\dot{y}_{11}, \quad -2\nu\psi_{21} = 2y_{11}. \tag{3.65}$$

Pour que le système (3.65) admette une solution il est nécessaire et suffisant que $y_{11}$ soit solution de l'équation

$$\frac{\dot{y}_{11}}{y_{11}} = -\frac{b^2}{2\nu},$$

d'où

$$y_{11} = C_1 \exp\left\{-\frac{1}{2}\int_0^t \frac{b^2}{\nu}\, dt\right\}. \tag{3.66}$$

Posons maintenant $\omega = \omega_2 = +\nu$. Comme dans ce cas $\omega - \nu = 0$, on déduit de la deuxième équation du système (3.62) que la fonction $\psi_1 = \psi_{12}$ est arbitraire. La première équation (3.62) nous permet alors d'exprimer $y_{12}$ en fonction de $\psi_{12}$:

$$y_{12} = \frac{b^2}{4\nu}\,\psi_{12}. \tag{3.67}$$

On déterminera $\psi_{12}$ à partir de l'équation (3.63) qui s'écrit ici:

$$2\nu y_{22} - \frac{1}{2} b^2 \psi_{22} = -\frac{d}{dt}\left(\frac{b^2}{4\nu}\,\psi_{12}\right), \quad -\dot{\psi}_{12} + \frac{b^2}{2\nu}\,\psi_{12} = 0. \tag{3.68}$$

Pour que le système (3.68) admette une solution, il faut et il suffit que

$$\dot{\psi}_{12} = \frac{b^2}{2\nu}\,\psi_{12},$$

d'où

$$\psi_{12} = C_2 \exp\left\{\frac{1}{2}\int_0^t \frac{b^2}{\nu}\, dt\right\}, \tag{3.69}$$

et en vertu de (3.67)

$$y_{12} = \frac{C_2 b^2}{4\nu} \exp\left\{\frac{1}{2}\int_0^t \frac{b^2}{\nu}\, dt\right\}. \tag{3.70}$$

Cherchons maintenant une intégrale particulière du système (3.59) à l'aide des méthodes envisagées au chapitre V. Représentons une solution particulière par des séries des puissances négatives de $\lambda$:

$$\tilde{y} = y^0 + \lambda^{-1} y^1 + \lambda^{-2} y^2 + \ldots,$$
$$\tilde{\psi} = \psi^0 + \lambda^{-1}\psi^1 + \lambda^{-2}\psi^2 + \ldots$$

En portant ces séries dans le système (3.59), on peut trouver des fonctions $y^i$, $\psi^i$. Signalons toutefois que pour trouver les premiers termes de ces développements, il nous suffit d'égaler à zéro les seconds membres du système (3.59):

$$\begin{aligned} -\lambda\nu y^0 + \frac{1}{2}\lambda b^2 \psi^0 + \lambda f &= 0, \\ \lambda\nu\psi^0 + 2y^0 &= 0. \end{aligned} \tag{3.71}$$

La résolution du système (3.71) nous donne

$$y^0 = \frac{\nu f \lambda}{\lambda \nu^2 + b^2}, \qquad \psi^0 = -\frac{2f}{\lambda \nu^2 + b^2}. \tag{3.72}$$

En se servant des expressions de $\psi_{11}, \psi_{12}, y_{11}, y_{22}, \tilde{y}$ et $\tilde{\psi}$ on obtient l'intégrale générale du système (3.59) sous la forme

$$y = C_1 \exp\left\{-\lambda \int_0^t \nu\, dt - \frac{1}{2} \int_0^t \frac{b^2}{\nu}\, dt\right\} + \\ + C_2 \exp\left\{\lambda \int_0^t \nu\, dt + \frac{1}{2} \int_0^t \frac{b^2}{\nu}\, dt\right\} \frac{b^2}{4\nu} + \tilde{y}, \tag{3.73}$$

$$\psi = C_2 \exp\left\{\lambda \int_0^t \nu\, dt + \frac{1}{2} \int_0^t \frac{b^2}{\nu}\, dt\right\} + \tilde{\psi}.$$

On détermine les constantes $C_1$ et $C_2$ à l'aide des conditions aux limites. Considérons maintenant le cas où les états initial et final du système sont fixes:

$$y\,(0) = y_0, \qquad y\,(T) = y_T. \tag{3.74}$$

En portant les formules (3.73) dans les conditions (3.74), on obtient les équations suivantes pour $C_1$ et $C_2$:

$$\begin{gathered} C_1 + C_2 = y_0 - \tilde{y}(0), \\ C_1 \exp\left\{-\lambda \int_0^T \nu\, dt - \frac{1}{2} \int_0^T \frac{b^2}{\nu}\, dt\right\} + \\ + C_2 \exp\left\{\lambda \int_0^T \nu\, dt + \frac{1}{2} \int_0^T \frac{b^2}{\nu}\, dt\right\} \frac{b^2}{4\nu} = y_T - \tilde{y}\,(T). \end{gathered} \tag{3.75}$$

La solution approchée du système (3.75) peut être exprimée sous une forme compendieuse. En résolvant ce système par rapport à $C_1$ et $C_2$ et en négligeant les termes $\exp\left\{-\lambda \int_0^T \nu\, dt\right\}$, on obtient sans peine

$$\begin{gathered} C_1 \sim y_0 - \tilde{y}\,(0), \\ C_2 \sim [y_T - \tilde{y}\,(T)] \exp\left\{-\lambda \int_0^T \nu\, dt - \frac{1}{2} \int_0^T \frac{b^2}{\nu}\, dt\right\} \frac{4\nu}{b^2}. \end{gathered} \tag{3.76}$$

Mettons maintenant l'intégrale générale (3.73) sous la forme:

$$y = (y_0 - \tilde{y}(0)) \exp\left\{-\lambda \int_0^t \nu\, dt - \frac{1}{2}\int_0^t \frac{b^2}{\nu}\, dt\right\} + [y_T - \tilde{y}(T)] + \tilde{y},$$

$$\psi = [y_T - \tilde{y}(T)]\, \frac{4\nu}{b^2} + \tilde{\psi}. \tag{3.77}$$

Une fois qu'on a trouvé $y$ et $\psi$, on détermine la commande optimale au moyen de la formule (3.57). Ce que nous voulions.

Nous avons résolu ce problème en détail pour illustrer les possibilités offertes par les formules asymptotiques.

Attirons, en conclusion, l'attention du lecteur sur une autre particularité des problèmes envisagés. La résolution implique l'analyse des valeurs propres de la matrice $\Gamma - \omega E$, où $E$ est la matrice unité, et

$$\Gamma = \begin{pmatrix} A & B \\ 0 & -A^* \end{pmatrix}.$$

Il est immédiat de prouver que l'ensemble des racines de l'équation caractéristique

$$|\,\Gamma - \omega E\,| = 0 \tag{3.78}$$

est la réunion des ensembles des racines des équations

$$|\,A - \omega E\,| = 0 \text{ et } |\,A + \omega E\,| = 0.$$

Une particularité du spectre de la matrice $\Gamma$ est que si $\omega = \mu$ est racine de l'équation $|\,A - \omega E\,| = 0$, donc de l'équation (3.78), alors $\omega = -\mu$ sera aussi racine de l'équation (3.78). Si l'équation $|\,A - \omega E\,| = 0$ admet des racines simples $\mu_i$ telles que

$$\mu_i = -\mu_j, \qquad i \neq j, \tag{3.79}$$

alors on peut appliquer la technique qui vient d'être développée pour construire les solutions approchées. Si parmi les racines de l'équation $|\,A - \omega E\,| = 0$, il en existe deux qui vérifient la relation (3.79), alors le problème se complique singulièrement: l'équation (3.78) admet alors des racines multiples. La procédure proposée dans ce numéro est inadéquate: les représentations asymptotiques devront contenir $\lambda$ à des puissances fractionnaires. Quelques considérations sur les méthodes de construction de telles représentations ont été exposées à la fin du chapitre précédent.

La condition (3.79) est souvent remplie. Si par exemple le système est conservatif, alors elle l'est toujours. Cette condition a lieu pour une bien plus vaste classe de systèmes. En effet, pour qu'elle soit réalisée, il suffit que le système possède au moins un élément oscillant, c'est-à-dire que l'équation $|\,A - \omega E\,| = 0$ admette deux racines imaginaires pures: $\omega_1 = i\mu$, $\omega_2 = -i\mu$. Donc le cas où les représentations asymptotiques sont des séries de puissances fractionnaires de $\lambda$ n'est pas si rare.

## CHAPITRE VII

# EXPERTISES ET PROCÉDURES NON FORMELLES

### § 1. Remarques préliminaires

Dans les chapitres précédents nous avons à plusieurs reprises attiré l'attention sur le fait que la résolution des problèmes compliqués nous incitait à recourir à des méthodes non formelles et que les problèmes d'analyse d'un système ou de son projet n'étaient pas tous justiciables d'une position mathématique satisfaisante. L'étude d'un système ou la qualité d'un projet dépendent pour beaucoup de l'habileté de l'analyste à inclure des méthodes mathématiques formelles dans la procédure non formelle d'analyse.

Bien plus, on a souligné que l'une des principales vocations de l'analyse des systèmes était d'apprendre à combiner les méthodes d'analyse mathématiques et les méthodes non formelles, les méthodes rigoureuses d'analyse des modèles formalisés et les expériences ou les devis des experts.

Mais est-il légitime d'envisager le principe d'une combinaison des méthodes mathématiques et des méthodes fondées sur l'intuition et l'expérience? La question posée a-t-elle un sens?

En effet, dans les mathématiques traditionnelles, seuls les résultats énoncés sous forme de théorèmes sont indubitables, eux seuls nous semblent crédibles. Les méthodes mathématiques et elles seules permettent d'aboutir à des conclusions irréfutables et non ambiguës. Les procédés d'analyse fondés sur l'intuition et les similitudes ne sont pas caractérisés par la même rigueur des conclusions. Toute proposition non acquise par des méthodes mathématiques peut être mise en doute par le mathématicien qui est toujours en mal de démonstrations. Cela tient de la déformation professionnelle.

Mais cette « déformation » qui occulte l'analyse des données initiales pour s'attacher seulement à celles des conséquences restreint le champ d'activités et les possibilités des mathématiques modernes et des mathématiciens dont le rôle ne cesse de croître dans la vie sociale. Elle éloigne le mathématicien de problèmes dont sans doute lui seul serait en mesure de faciliter l'analyse.

Il existe cependant un autre point de vue sur les objectifs et la teneur des recherches mathématiques. La tendance qui se dessine actuellement peut tenir en une phrase: le mathématicien devient de plus en plus un membre actif de l'analyse de processus biologiques, physiques, économiques, etc. On assiste en quelque sorte à un retour

25—01523

à l'époque de la Renaissance où chaque grand mathématicien était encore un philosophe, un naturaliste. Mais ce retour s'opère au stade actuel de nos connaissances sur l'environnement et des possibilités accrues des recherches mathématiques.

A noter qu'en mathématiques le formel et le non formel se côtoient et il est parfois très difficile de faire la distinction entre la partie euristique de l'analyse qui est fondée sur l'intuition et l'étude de l'environnement, et les constructions mathématiques formelles. En effet, les mathématiques tirent des conclusions rigoureuses à partir des données initiales. Mais ces données — les axiomes — découlent d'hypothèses paramathématiques. Ces hypothèses résultent d'un raisonnement non formel, d'une extrapolation de l'expérience et des observations. Donc, le formel et le non formel s'imbriquent dans les recherches *). Le modèle mathématique est le fruit d'un raisonnement non formel; toute l'information concernant la nature du processus étudié y est codée. On construit ensuite une algèbre, c'est-à-dire un système de procédures dont l'algorithme permet de décoder l'information logée dans le modèle. Donc, l'une des tâches des mathématiques est de décoder l'information contenue dans le modèle, de construire une séquence d'opérations logiques que la manière non formelle de raisonnement, qui est traditionnelle pour les sciences naturelles, est incapable de faire **).

Mais si les méthodes formelles et non formelles d'analyse s'imbriquent si étroitement, il semble tout à fait naturel de ne pas les désagréger et de les considérer comme les éléments d'un processus unique de recherche. Donc, on peut ne pas dissocier la mise en place d'un système d'hypothèses (d'axiomes), c'est-à-dire la conception des modèles, des méthodes d'étude de ce système. Nous verrons que ce principe donnera lieu à des recommandations pratiques bien définies.

La description mathématique, c'est-à-dire la construction d'un modèle mathématique, n'est pas un processus univoque. S'il est vrai que le modèle est objectif, l'activité de l'analyste fait jouer d'innombrables facteurs subjectifs: minutie de la description du processus, choix du langage de simulation, etc. La construction d'un modèle mathématique s'appuie sur un système d'hypothèses qui reflètent la vision de l'analyste. A cela il faut ajouter encore le degré de fidélité du modèle. Nous avons abordé ces sujets au chapitre III lorsque nous avons parlé de l'aspect informationnel du problème. L'existence de « degrés de liberté » dans la description du processus étudié permet de construire des modèles correspondant aux

*) Ce sujet a été magistralement traité par le mathématicien italien G. Polya [61].

**) Pour être plus précis, elle peut le faire mais pour cela elle doit être formalisée (formellement décrite dans un langage quelconque sous forme de procédures canoniques).

possibilités de l'appareil, c'est-à-dire aux possibilités des méthodes mathématiques et des ordinateurs mis entre les mains de l'analyste. Donc la construction du modèle mathématique fait corps avec le matériel d'analyse. Cette proposition est l'un des principaux axiomes de l'analyse des systèmes et elle contredit dans une certaine mesure le principe de séparation de ces deux composantes du processus d'analyse. En effet, au XIX$^e$ siècle déjà on avait une idée assez nette de l'existence de deux étapes indépendantes: la construction du modèle, c'est-à-dire la description mathématique du phénomène étudié, et l'étude du modèle. Ce point de vue a été clairement exprimé par A. Liapounov qui estimait qu'une fois posé, tout problème de mécanique ou de physique devait être traité comme un problème de mathématiques. Ce point de vue est partagé par l'écrasante majorité des mathématiciens. La vision développée au fil de cet ouvrage diffère du point de vue traditionnel: le mathématicien qui analyse un système ne doit jamais oublier l'objet et le contenu de son étude. Il faut dire que cette approche n'est pas nouvelle. Au siècle dernier déjà un savant de renom a déclaré: « Le bon mécanicien n'est pas celui qui a réussi à mettre un mouvement en équations, mais celui qui a trouvé des équations intégrables. » *) Cette boutade exprime en fait la même idée.

On traitera donc le processus d'analyse comme un seul processus regroupant les méthodes formelles et non formelles d'analyse. Mais une fois qu'on a souscrit à ce point de vue, on doit organiser d'une certaine manière le système de procédures que nous conviendrons d'appeler non formelles, ou euristiques. Le sens de ces procédures est d'universaliser les raisonnements, de les ordonnancer, bref, de les formaliser. En décrivant ces procédures il faudra toujours signaler les hypothèses utilisées, celles qui ne résultent pas d'autres hypothèses plus simples adoptées antérieurement. Une description correcte des hypothèses, leur séparation des futures constructions formelles donnent une idée de la crédibilité des résultats obtenus. Ceci est un autre principe important de l'analyse des systèmes.

Dans ce chapitre nous allons tenter de décrire quelques classes de procédures non formelles fréquemment utilisées en analyse des systèmes et d'illustrer ainsi les diverses possibilités qui se présentent ici (leur exposé systématique est impossible par manque d'expérience).

La nécessité d'une analyse non formelle se présente non seulement au niveau de la formulation des hypothèses, mais au niveau aussi de l'élaboration des méthodes mathématiques lorsque, par exemple, la solution exacte du problème existe mais sa réalisation implique une très longue occupation de l'ordinateur. Nous avons déjà eu affaire

*) De nombreux auteurs attribuent ces mots à N. Joukovski. Personnellement je n'ai pas réussi à trouver un document à l'appui.

à quelques méthodes semblables dans les chapitres précédents. De nombreux nouveaux problèmes et méthodes euristiques ont vu le jour depuis que les ordinateurs participent aux expériences. Au chapitre III nous avons, en décrivant la structure des systèmes de simulation, cité quelques exemples de procédures euristiques utilisées dans le dialogue homme — machine. L'apparition de puissants ordinateurs et de systèmes de simulation a fourni un instrument spécial de synthèse des méthodes formelles et non formelles, une synthèse qui, soulignons-le, a toujours existé. Les systèmes de simulation ont permis d'élaborer seulement de nouveaux types de procédures et sans plus (des exemples ont été donnés au chapitre III). Mais les méthodes de l'analyse des systèmes sollicitent largement les autres procédures euristiques de prise de décision et les informations recueillies à l'aide de l'intuition, l'expérience et les connaissances des analystes, constructeurs et autres que nous conviendrons d'appeler experts.

L'utilisation des experts, c'est-à-dire le recours direct de l'analyste non pas à l'ordinateur mais aux connaissances et à l'expérience de l'homme, ouvre encore une voie pour l'étude des procédures non formelles et leur application en analyse des systèmes. L'expert remplit en quelque sorte les fonctions d'un instrument qui soit procède à un choix, soit fixe les valeurs des coefficients, soit établit un lien logique entre la cause et l'effet, etc. Donc, le recours à l'expert est une sorte de test qu'il est naturel d'affiner, de répéter, de faire réaliser par plusieurs experts pour profiter de leur expérience collective.

L'usage de l'expérience collective, la préférence qui lui est accordée dans telle ou telle situation sont encore une hypothèse dont l'analyste porte l'entière responsabilité. L'utilisation des expertises collectives a conduit à divers modes de vote dont la description et l'analyse ont fait l'objet de nombreux travaux.

Comme exemples de procédures traditionnelles faisant appel à l'expérience collective, citons les expertises, les conseils, les délibérations, etc. Ces procédures peuvent fortement différer selon la nature de la situation. Dans certains cas, la discussion peut servir de base à ces procédures, dans d'autres, elle est catégoriquement exclue. Dans une consultation de médecins par exemple, il est très important de débattre toutes les opinions, de ne laisser aucun détail au hasard, de bien peser le pour et le contre, etc. Donc, le diagnostic est une procédure qui encourage la discussion.

Un autre exemple de délibérations est décrit par L. Tolstoï dans « Guerre et paix ». Au conseil qui se tint à Fili, le feld-maréchal Koutouzov demanda à chaque général convoqué son analyse de la situation et ses propositions. La subordination était strictement respectée : la parole était donnée aux moins gradés et ensuite aux plus gradés. Le dernier mot revenait au feld-maréchal. Après avoir écouté toutes les opinions, il déclara : « J'ordonne... » La décision était prise à l'« unanimité » et était sans appel.

Dans bien des cas il est tout simplement faux de postuler l'avantage des décisions collectives. Les opérations militaires nous fournissent d'innombrables exemples. Hannibal n'aurait probablement jamais battu les Romains à Cannes s'il s'était servi des méthodes de décision collective.

Signalons que l'avis de la majorité traduit des avis « moyens », donc il est nécessaire lorsque nous voulons connaître des caractéristiques moyennes. Trouver une solution exclusive par un sondage est une gageure. Dans de telles conditions, l'avis d'un homme éclairé vaut souvent bien plus qu'une expertise collective. Cette proposition n'est pas rigoureuse non plus : il est pratiquement impossible d'indiquer les conditions sous lesquelles l'opinion d'un expert est à préférer à un avis collectif. Ici aussi c'est l'expérience qui prime.

Le mode d'organisation des expertises et des délibérations est souvent réglementé par la tradition (en fin de compte par l'expérience) et relève souvent de l'art, c'est-à-dire est défini par un système d'hypothèses. Mais les méthodes mathématiques de traitement de l'information s'insinuent progressivement dans ce domaine.

L'un des premiers exemples d'application des « méthodes mathématiques » dans les expertises nous est donné par l'industrie vinicole. Les instituts de dégustation sont nés avec Bacchus. Leur objectif est d'apprécier la qualité d'un vin (au moyen d'une note), de classer les vins d'après leur qualité, etc. En attribuant la note moyenne d'un vin, les dégustateurs procèdent à un traitement élémentaire de l'information. Cette procédure se complique dès lors que l'on tient compte du « poids » des experts. En effet, le « poids » de l'expert dépend aussi bien de l'expérience de ce dernier que des résultats de son activité. Ainsi le « poids » de l'expert (la rémunération de l'expert dépend de son « poids ») est d'autant plus petit que l'écart de sa note par rapport à la moyenne est élevé. Cette complication de la procédure d'appréciation introduit un élément d'apprentissage. Le mode de fonctionnement d'une équipe d'experts se transforme en un processus dynamique à rétroaction, puisque la situation de chaque expert (c'est-à-dire son « poids ») dépend de ses activités professionnelles. Ces derniers temps, des tentatives sont entreprises pour compliquer davantage ces procédures. C'est ainsi qu'on essaye de prendre la demande en considération et d'adapter le « goût collectif » des dégustateurs à la conjoncture.

L'exemple des dégustateurs est un exemple classique d'expertises simples qui sont couramment utilisées pour régler des problèmes relativement peu compliqués. Il existe actuellement des méthodes variées d'organisation d'expertises et d'établissement de devis.

Les instructions diverses destinées à vérifier le bon fonctionnement des systèmes techniques complexes nous fournissent un autre exemple d'expertises simples. Les instructions indiquent toujours exactement par quoi il faut commencer l'inspection. Si une panne

est détectée, l'instruction indique les observations supplémentaires à faire, les mesures préventives à prendre ou enfin le moyen de réparer l'élément défectueux. La personne qui suit les indications de l'instruction n'est pas censée comprendre la nature de la panne. Tout ce qu'on lui demande, c'est de remettre le système en état de marche. Un système informatique ou un réseau énergétique sont si compliqués qu'il est totalement exclu que le personnel exploitant en connaisse tous les détails. L'instruction est généralement composée par une équipe de « superspécialistes » compte tenu de l'expérience d'exploitation de systèmes identiques ou semblables. L'essentiel est que l'instruction est constamment mise à jour grâce à l'information reçue par les experts.

Les instructions sont aussi des expertises et leur champ d'action ne s'étend pas aux seuls systèmes techniques. Traitées comme un service, elles peuvent manifestement être utilisées dans la gestion économique, voire en médecine. En tout cas leur implantation aurait permis de réduire considérablement le temps nécessité par des analyses souvent inutiles et par la prescription d'un traitement, car tout médecin bénéficierait des tout derniers acquis.

Ce qui vient d'être dit semble couler de source et pourtant les idées liées aux « instructions de détection des pannes » se fraient difficilement la voie, notamment en médecine. La cause n'est pas uniquement dans les efforts impliqués par la composition et la mise à jour de telles instructions, elle est aussi dans les réactions conformistes des exécutants et des usagers.

## § 2. Exemples d'expertises complexes

L'exemple des dégustateurs fait partie des expertises simples dans lesquelles chaque expert est en mesure de donner une réponse qualifiée à la question posée. Cette question peut être assez compliquée. L'essentiel, c'est que l'expert soit capable d'y répondre, que sa compétence rende ses conclusions crédibles, assez crédibles pour aider l'analyste à prendre des décisions. Il existe des problèmes pour lesquels il est impossible de trouver des experts capables de fournir une appréciation valable. Et pourtant même dans ces cas on peut utiliser des expertises et des devis. Il suffit de conduire l'expertise d'une manière bien spéciale. Le problème doit être dûment apprêté, décomposé en une suite de sous-problèmes simples accessibles aux experts. Il faut en quelque sorte effectuer des expériences avec les experts. Dans ce paragraphe, on se propose d'étudier deux exemples illustrant la technique de décomposition dans les expertises.

a) *Méthode de l'arbre des objectifs.* La prévision des situations (scientifiques, techniques, politiques, etc.) par des appréciations d'expert est un problème très répandu. Si l'objet étudié est assez complexe, l'expert est souvent incapable de fournir une réponse

suffisamment claire à la question posée : sous cette forme la question dépasse sa compétence.

Supposons, par exemple, que la prévision de l'événement $A$ consiste à répondre à la question : « L'événement $A$ aura-t-il lieu durant un intervalle de temps $t \leqslant T$ ou non ? ». En général l'expert est incapable de donner une réponse précise à une telle question. Il parlera d'une réalisation « plus ou moins probable » de l'événement ou encore des chances de réalisation de cet événement, etc. C'est pourquoi cette question est posée généralement en termes probabilistes : « Quelle est, selon l'expert, la probabilité $P(T)$ que l'événement $A$ se produise durant un intervalle de temps $t \leqslant T$ ? ». Cette estimation sera dite probabilité intuitive. Signalons d'emblée que cette estimation a un rapport purement conventionnel avec la notion mathématique de probabilité, car il est question ici d'un seul événement. Nous verrons plus bas que cette forme de représentation non rigoureuse de l'information est très commode pour la résolution des problèmes pratiques.

REMARQUE. La prise d'une décision, je dis bien d'*une* décision, sur la base d'une analyse des probabilités intuitives doit être traitée comme une hypothèse, par exemple de la nature suivante : « l'utilisation des probabilités intuitives fournit dans la plupart des cas de bons résultats pratiques ».

La deuxième difficulté est liée au fait que les événements à prédire sont assez complexes et l'expert n'a pas la compétence requise pour fournir une réponse précise. Dans ce cas l'événement doit être désagrégé, c'est-à-dire qu'il faut construire un arbre d'événements élémentaires justiciables d'expertises du type signalé ci-dessus.

Supposons, par exemple, qu'il s'agisse de dire si l'homme mettra les pieds sur Mars en l'an 2000. Nous conviendrons d'appeler cet événement $S$, événement final. Chaque membre de l'équipe d'experts doit recenser les événements $S_1, S_2, \ldots, S_k$ qui conditionnent la réalisation de l'événement $S$. Dans notre exemple, l'événement $S_1$ peut représenter la construction d'un moteur doué des caractéristiques requises, l'événement $S_2$, la création d'un système de survie, l'événement $S_3$, l'existence d'une station orbitale pour le montage d'appareils, etc. Définissons encore l'événement $S^f$ :

$$S^f = f(S_1, S_2, \ldots, S_k),$$

où $f$ est une fonction logique des variables $S_i$. Dans le cas le plus élémentaire, $f$ est formée à l'aide exclusivement de conjonctions, c'est-à-dire que l'événement $S^f$ consiste en l'apparition simultanée des événements $S_1, S_2, \ldots, S_k$. La fonction $f$ peut être bien plus complexe. Si, par exemple, le vaisseau peut être assemblé ailleurs que sur une station orbitale, alors l'opération ET peut être remplacée par l'opération OU.

Pour analyser de telles situations V. Glouchkov [30] a proposé un procédé ramenant le problème de l'expertise à l'évaluation de la

probabilité conditionnelle $P^f(t)$ que l'événement $S$ se produise sur un intervalle de temps $\leqslant t$, sachant que la fonction $f$ prend une valeur égale à l'unité, c'est-à-dire que l'événement $S^f$ s'est réalisé. Ceci constitue l'étape finale de l'expertise. Dans les étapes précédentes, la détermination de la probabilité de l'événement $S^f$ est généralement confiée à d'autres experts.

Supposons tout d'abord que les experts peuvent désigner les probabilités $P_i(\tau)$ de réalisation des événements $S_i$ sur un intervalle de temps $t \leqslant \tau$. La procédure d'expertise s'arrête alors là et l'on peut passer au traitement de ses résultats. Illustrons ceci sur l'exemple élémentaire de la fonction $f$. Vu que l'événement $S^f$ consiste en la réalisation simultanée des événements $S_i$, la probabilité $P_f(\tau)$ qu'il se produise sur un intervalle de temps $t \leqslant \tau$ sachant que les événements $S_i$ sont indépendants, sera égale à

$$P_f(\tau) = P_1(\tau) P_2(\tau) \ldots P_k(\tau).$$

La principale tâche des experts, comme nous l'avons déjà dit, consiste à déterminer la probabilité conditionnelle de la réalisation de l'événement $S$ sur un intervalle de temps $t$ après la réalisation de l'événement $S^f$. On obtient cette fonction en moyennisant (compte tenu ou non du « poids » des experts) les appréciations des experts. Désignons-la par $Q(t)$. Désormais il est immédiat de calculer la probabilité de réalisation de l'événement $S$ sur un intervalle de temps $t \leqslant t^*$:

$$P(t) = \int_0^{t^*} Q(t-\xi)\, dP_f(\xi). \tag{2.1}$$

Si les experts ne peuvent pas définir la probabilité $P_i(\tau)$ de réalisation de l'événement $S_i$, autrement dit si l'événement $S_i$ est trop compliqué, il faut alors passer à l'étape suivante, laquelle consiste à indiquer les événements $\{S_{ij}\}$ qui conditionnent l'apparition de l'événement $S_i$, et ensuite à introduire l'événement $S_i^f$ qui consiste en la réalisation des événements $S_{ij}$. Si les experts sont capables d'évaluer les probabilités $P_{ij}(\tau)$ de réalisation des événements $S_{ij}$ sur un intervalle de temps $t \leqslant \tau$ et la probabilité conditionnelle $Q_i(T)$ de réalisation de l'événement $S_i$ sur un intervalle de temps $T$ après la réalisation de l'événement $S_i^f$, alors une formule de la forme (2.1) nous donnera la probabilité de réalisation de l'événement $S_i$.

Si les experts ne peuvent pas évaluer les probabilités $P_{ij}(\tau)$, alors on poursuit le processus de désagrégation. Ce qui nous amène à un arbre d'événements (fig. 7.1). En continuant cette procédure on finira par obtenir des événements assez simples à évaluer par les experts.

Ce procédé de raisonnements a fait fortune en analyse des systèmes. L'exemple des estimations probabilistes est l'une des innombra-

bles illustrations de la technique de traitement des expertises, technique dite d'utilisation de l'arbre des objectifs. Cette technique est largement utilisée en planification, dans la conception des grands projets, etc. L'idée maîtresse du schéma décrit, de même que des autres variantes de la méthode de l'arbre des objectifs, est la dé-

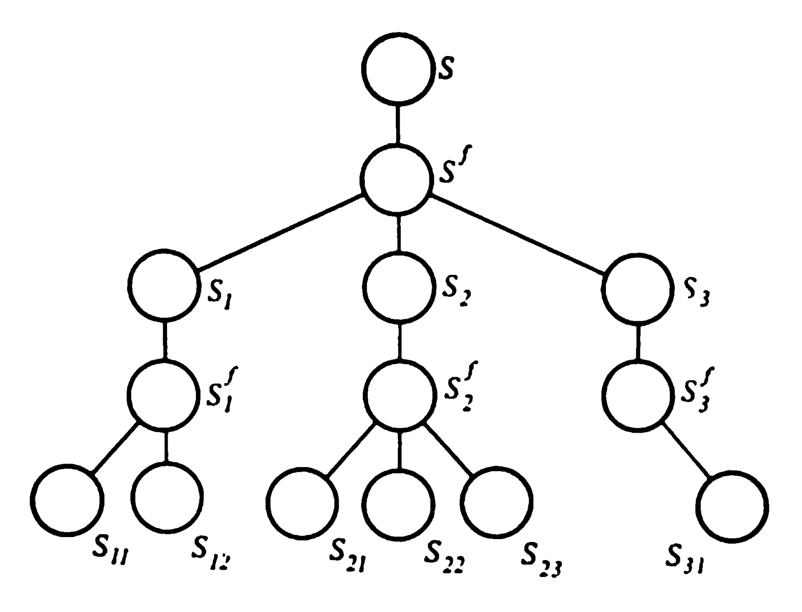



Fig. 7.1

sagrégation d'un gros problème en sous-problèmes. Pour ce qui est des estimations, elles peuvent aussi bien porter sur les probabilités que sur les coûts, la fiabilité, etc.

REMARQUE. L'arbre des objectifs a une valeur intrinsèque en analyse des systèmes. C'est un schéma architectural du projet ; quant à la représentation de l'information sur le projet sous la forme d'un arbre des objectifs, elle constitue un maillon important de la technique de projection par sa suggestivité.

Ainsi, par « arbre des objectifs » ou « méthode de l'arbre des objectifs » on entendra un procédé de désagrégation d'un problème en sous-problèmes, la représentation d'un projet, qu'il soit physique, technique ou économique, sous la forme d'un graphe.

La méthode de l'arbre des objectifs combinée aux procédures d'expertise est l'une des principales méthodes de désagrégation. Aux probabilités fournies par les experts, on peut substituer des modèles mathématiques et des estimations acquises par des méthodes d'analyse formalisées. Sous cette forme la méthode de l'arbre des objectifs se transforme en un outil assez universel de prédiction. On peut l'utiliser, par exemple, pour l'analyse de situations internationales, politiques, économiques, militaires. Certes, chaque cas implique une procédure spéciale, mais le principe d'analyse est en gros le même.

La méthode de l'arbre des objectifs n'est pas l'unique méthode de décomposition et de plus elle ne s'applique qu'à une certaine classe de problèmes.

b) *Méthode des matrices résolvantes.* Ainsi, le succès d'une expertise complexe dépend essentiellement de notre aptitude à décomposer le problème envisagé en sous-problèmes. Il n'y a pas de panacée ici. La méthode de l'arbre des objectifs est une méthode de décomposition parmi beaucoup d'autres. G. Pospelov a proposé une méthode pour une classe de problèmes auxquels on ne peut appliquer les raisonnements ci-dessus (voir [62]). Illustrons la méthode des matrices résolvantes sur le problème de ventilation des ressources destinées à la recherche scientifique. A l'époque c'était un véritable problème : comment ventiler le budget de la recherche scientifique et réduire à son plus bas niveau le facteur subjectif dans la résolution de ce problème capital. Ce problème est par essence difficile, car l'enquête effectuée directement auprès des chercheurs intéressés ne fournit pratiquement aucune information utile à l'élaboration d'un principe de ventilation des ressources. Le planificateur du budget doit trouver un juste milieu entre les désirs naturels du savant pour qui la recherche est capitale et les intérêts de la société qui ne peut consacrer que des sommes limitées à la recherche. Les intérêts de la société doivent être représentés par un programme sous forme de liste des objectifs. C'est un point très important : si les objectifs du développement de la recherche scientifique ne sont pas clairement définis, il est complètement farfelu de comparer les divers principes de ventilation des ressources. Ce problème est de ceux auxquels il faut appliquer la méthode de programmation. Il n'existe pas d'autre alternative dans cette situation. Voyons donc les objectifs. Ils peuvent être très variés. Certains seront scientifiques : envoi d'hommes sur Mars en l'an 2000 ou création d'un laboratoire sous-marin aux larges des îles Kouriles. D'autres, économiques : construction d'une centrale thermonucléaire ou robotisation de tel ou tel processus technologique. D'autres, enfin, seront militaires : installation d'un système de défense anti-missile, etc.

Nous escamotons le problème de l'élaboration du programme. En fait, c'est aussi une procédure définie par l'orientation imprimée au développement du potentiel scientifique d'un pays. Le potentiel scientifique n'a pas une valeur intrinsèque. Les objectifs scientifiques sont la conséquence d'autres objectifs plus généraux : politiques, économiques, militaires, etc. Nous ne discutons pas ici les liens des objectifs scientifiques et des objectifs généraux et nous admettrons que la liste des travaux (des objectifs scientifiques) est donnée. Désignons-la par $\alpha$, vecteur de composantes $\alpha^i$ ($i = 1, 2, \ldots, n_\alpha$).

La procédure de ventilation des ressources suppose que toutes les composantes du vecteur $\alpha$ sont pondérées, c'est-à-dire que des valeurs strictement positives bien définies sont attribuées à ces com-

posantes. Ces valeurs doivent être normées d'une manière ou d'une autre: on peut par exemple prendre $\sum_{i=1}^{n_\alpha} \alpha^i = 100$ ou $\sum_{i=1}^{n_\alpha} \alpha^i = 1$.

La suite de la procédure sera calquée sur le mode d'organisation de la recherche dans notre pays, c'est-à-dire qu'elle comportera les trois étapes suivantes: construction et manipulation du modèle, recherches appliquées, recherches théoriques. Donc, l'étape suivante consiste à établir la liste des travaux exigés par la construction et la manipulation du modèle. Une équipe d'experts dresse une liste $\beta$ (un vecteur de composantes $\beta^i$) de travaux censés réaliser les objectifs fixés. Mais on ne peut exiger des experts qu'ils désignent les poids des coordonnées $\beta^i$. En effet, sur quel critère peut-on comparer la construction d'un moteur pour un véhicule spatial et la construction d'un robot? Seulement en établissant une relation entre la liste $\beta$, la liste $\alpha$ et les poids des travaux de la liste $\alpha$. La procédure suivante a été proposée (cf. [62]) pour la détermination des poids de ces travaux. La construction et manipulation du modèle poursuit en principe plusieurs objectifs scientifiques. Donc, l'expert doit résoudre un problème relativement simple qui consiste à composer la matrice $A^\beta = (a_{ij}^\beta)$ des valeurs des travaux de la liste $\beta$. L'élément $a_{ij}^\beta$ est un nombre strictement positif désignant la valeur relative du travail $i$ pour l'objectif $j$. Ces quantités sont ensuite normées selon un procédé quelconque. On peut, par exemple, prendre $\sum_j a_{ij}^\beta = 1$. Pour poids du travail $i$, il est désormais naturel de prendre la quantité $\beta^i = \sum_j a_{ij}\alpha^j$. On obtient ainsi la formule

$$\beta = A^\beta \alpha, \qquad (2.2)$$

qui définit une application de l'ensemble des objectifs scientifiques sur l'ensemble des travaux nécessités par la construction et la manipulation du modèle.

Pour construire et manipuler le modèle il faut procéder à des recherches appliquées. Désignons la liste de ces recherches par le vecteur $\gamma$. Le problème des experts est maintenant de dresser la liste $\gamma$ et de composer la matrice $A^\gamma$ des valeurs des recherches appliquées.

Les experts qui composent les matrices $A^\beta$ et $A^\gamma$ ne sont pas en général les mêmes, puisque ces problèmes impliquent un niveau de qualification et une spécialisation différents. Une fois définie la matrice $A^\gamma$, on peut par les mêmes raisonnements construire une application de l'ensemble des travaux nécessités par la construction et la manipulation du modèle sur l'ensemble des recherches appliquées:

$$\gamma = A^\gamma \beta. \qquad (2.3)$$

La troisième étape enfin consiste en la définition, par un autre groupe d'experts, des orientations des recherches fondamentales. Par exemple, $\delta^1$ représentent les méthodes numériques en théorie des écoulements turbulents, $\delta^2$, les conditions de stabilité des populations en présence d'une radiation élevée, $\delta^3$, la classification des particules élémentaires, etc.

Ensuite le même groupe d'experts (aidés éventuellement par des spécialistes en recherches appliquées) compose la matrice $A^\delta$ des valeurs des recherches fondamentales nécessaires à la réalisation des recherches appliquées. Ceci nous permet de construire une application de l'ensemble des recherches appliquées sur celui des recherches théoriques:

$$\delta = A^\delta \gamma. \tag{2.4}$$

Les formules (2.2), (2.3) et (2.4) permettent de construire une application de l'ensemble des objectifs sur celui des recherches fondamentales:

$$\delta = A^\delta A^\gamma A^\beta \alpha. \tag{2.5}$$

Si les poids des composantes $\alpha^i$ sont donnés, on peut calculer les poids des composantes du vecteur $\delta$.

Cette procédure définit le poids des diverses recherches fondamentales dans la réalisation du programme envisagé. Le planificateur du budget a maintenant beau jeu de ventiler les ressources consacrées aux recherches fondamentales: il doit les répartir proportionnellement aux poids des composantes du vecteur $\delta$.

REMARQUE. La ventilation du budget est un problème assez complexe même si les composantes du vecteur $\delta$ sont déterminées. Parfois il est insensé d'entamer certaines recherches, par exemple si le budget ne suffit même pas à couvrir les frais d'achat des équipen. nts. Par ailleurs il ne faut pas interrompre certaines recherches même si leur po.ds est nul. Quoi qu'il en soit, l'information fournie par la détermination des poids du vecteur $\delta$ est une base rationnelle pour résoudre le problème de la ventilation du budget de la recherche scientifique.

La méthode des matrices résolvantes qui fait dépuis longtemps partie de l'arsenal des méthodes de l'analyse des systèmes est appliquée à la résolution de divers problèmes techniques et économiques.

Au chapitre III nous avons exposé les grandes lignes de la méthode de programmation. Nous l'avons définie comme un système de procédures rattachant les thèses d'une doctrine — les directives du Parti — à un train de mesures économiques nécessaires à la réalisation de ces thèses et aux instruments réalisant ces mesures. Donc, le train de mesures et la ventilation des ressources nécessaires à sa mise en œuvre sont des éléments très importants de la réalisation de la méthode de programmation. L'organisation des recherches scientifiques fondamentales fait partie de ces mesures. Donc, la liste $\alpha$ des objectifs doit résulter, comme déjà signalé, des directives, et leurs poids, que nous avons supposé donnés, peuvent être en réalité

déduits à partir des objectifs généraux du pays par la méthode décrite. A cet effet il nous faut ajouter un « étage » au schéma de résolution du problème de ventilation du budget des recherches fondamentales. A la lumière de ce qui précède on mesure toute l'importance que revêt la méthode des matrices résolvantes pour l'affinement des principes de la gestion programmée. La méthode des matrices résolvantes s'applique à une grande classe de problèmes et notamment aux problèmes de développement d'une région.

Le plan général de développement d'une région commence par un recensement des objectifs. Une partie d'entre eux est exogène, l'autre est déterminée par les besoins intérieurs de la région. A prendre, par exemple, le bassin de l'Enisséi qui couvre la région de Krassnoïarsk, la république autonome de Touva et la région d'Irkoutsk. Parmi les objectifs exogènes on peut citer la création d'une infrastructure hydroénergétique et d'un système de centrales thermiques, la production des métaux non ferreux, etc. Les intérêts de la région sont : le maintien de l'équilibre écologique, l'amélioration du niveau de vie de la population (ravitaillement, logement, etc.). Ces directives constituent la liste $\alpha$. Pour les réaliser il faut résoudre de nombreux problèmes auxiliaires : création d'une infrastructure routière, accroissement de la production agricole dans le bassin de Minoussinsk, problème de la démographie, etc. Tout ceci forme la liste $\beta$. Tout travail de la liste $\beta$ conditionne de nombreux travaux de la liste $\alpha$.

Mais nous sommes encore loin du plan général de développement. Le premier pas consiste à dresser la liste des mesures. La méthode des matrices résolvantes ou de sa modification peut être d'une grande utilité, car elle permet non seulement d'ordonnancer les travaux mais aussi de les pondérer. L'étape suivante consiste à détailler les travaux sous la forme d'un graphe et à établir un calendrier des travaux d'après les données dont on dispose sur les ressources consacrées à leur exécution.

c) *Discussion et commentaires.* Malgré leur différence fondamentale, la méthode de l'arbre des objectifs et la méthode des matrices résolvantes partent d'un même principe : la décomposition des problèmes complexes en sous-problèmes plus simples. L'homme n'est capable d'analyser une situation que si le nombre de facteurs (de corrélations) entrant en jeu est relativement peu élevé. Donc, l'avis de l'expert ne sera plus ou moins crédible que s'il aura à répondre à des questions relativement simples.

Si l'on avait demandé en 1959 à des experts d'évaluer la probabilité de débarquement de l'homme sur la Lune en 1969, leurs réponses auraient probablement été très loin de la vérité. D'autre part, si l'on avait préalablement décomposé le problème, par exemple désigné les principales étapes de la réalisation du programme « Apollo », la méthode exposée au début du paragraphe aurait donné une probabilité légèrement inférieure à l'unité.

Le succès d'une expertise dépend directement de la manière dont le problème est décomposé. En effet, le principal mérite de la méthode de la répartition du budget est surtout d'avoir savamment décomposé le problème. On ne peut exiger de l'expert qu'il donne un avis aussi éclairé sur la valeur relative des recherches en génétique radiative, la classification des particules élémentaires ou la création d'une théorie mathématique de la stabilité du mouvement d'un gaz. Cependant de nombreux experts apprécieront sans peine le rôle de l'information recueillie par des recherches fondamentales portant sur des problèmes appliqués.

Signalons qu'il n'existe aucune méthode universelle de décomposition des problèmes. Cette procédure dépend essentiellement de la nature du problème et de la compétence des analystes: la décomposition d'un problème complexe en sous-problèmes plus simples est une procédure euristique qui implique du savoir et de l'ingéniosité.

La formalisation des expertises qui a été décrite ne contredit pas le principe général des systèmes de simulation, développé au chapitre III, elle le prolonge. D'autre part, comme déjà signalé, l'organisation du dialogue homme — machine est un système de procédures de rétrécissement progressif de l'ensemble des décisions. Les raisonnements produits peuvent précisément être appliqués à cet objectif: ils permettent tout d'abord de rejeter les solutions *a priori* mauvaises et donc de retenir celles qui seront ultérieurement simulées.

## § 3. Méthodes euristiques dans les problèmes discrets

Dans les paragraphes précédents, on a parlé de l'usage des procédures non formelles dans les problèmes non formalisés. L'application des méthodes de décomposition du problème primitif peut être traitée comme une étape de formalisation de ce problème. On obtient ainsi des algorithmes dans la réalisation desquels l'expert doit être considéré comme un opérateur agissant sur le corps d'une certaine information.

Certes, nous sommes encore loin d'une véritable formalisation (du point de vue des mathématiques traditionnelles). Mais l'interprétation des procédures euristiques comme un algorithme avec des opérateurs humains (c'est-à-dire que l'opérateur est un homme) est très fructueuse. Elle permet de donner une description cohérente des diverses procédures homme — machine.

Jusqu'ici nous avons parlé de l'application des procédures euristiques à l'analyse des situations non formalisées et où les problèmes mathématiques n'étaient pas rigoureusement posés. Il est possible que les procédures euristiques soient nécessaires aussi dans les problèmes qui non seulement sont bien posés, mais qui de plus ne soulèvent pas de difficultés sur le plan mathématique, donc ne pré-

sentent pas un intérêt particulier du point de vue des mathématiques traditionnelles. C'est le cas notamment des problèmes impliquant un tri complet d'un nombre fini de variantes, c'est-à-dire des problèmes dont la solution existe toujours et dont la méthode de résolution est évidente, mais le temps nécessaire pour la trouver est si long que son utilisation perd tout intérêt. Etablir l'existence de la solution d'un problème est une chose, réaliser l'algorithme de résolution en est une autre. Nous sommes souvent contraints de recourir à des méthodes euristiques de calcul. Mais les méthodes euristiques ne sont « bonnes » que pour une classe spéciale de problèmes que seuls les experts peuvent signaler. Cette situation se présente très fréquemment dans les problèmes d'analyse et de projection de systèmes. Dans ce paragraphe, on se propose d'étudier quelques problèmes assez répandus portant sur la composition de calendriers de travaux.

a) *Composition d'un calendrier de travaux.* La composition des calendriers de travaux tient une place très importante dans la planification et la projection.

Nous avons déjà évoqué ce problème au chapitre premier. Rappelons-en les principaux traits. Supposons qu'on ait à effectuer une liste de travaux $P_1, P_2, \ldots, P_N$. L'exécution de ces travaux est soumise à certaines conditions que l'on peut classer en deux groupes.

*Condition* ($\alpha$). Les travaux doivent se succéder dans un ordre bien défini. Le travail $P_i$ ne peut commencer avant la fin d'une liste de travaux $P_{i1}, \ldots, P_{ik}$. La condition ($\alpha$) est une condition « logique »: elle rappelle un graphe orienté sans boucle.

*Condition* ($\beta$). C'est une condition de type ressources. Désignons par $v_i(t)$ le flot du vecteur ressources destiné à la réalisation du travail $P_i$. Les ressources étant limitées, on a

$$\sum_{i=1}^{N} v_i(t) \leqslant v(t). \tag{3.1}$$

Les contraintes de type intégrale

$$\int_0^T v^j(t)\,dt \leqslant V^j, \tag{3.2}$$

où $v^j(t)$ est le flot des ressources de la composante $j$ du vecteur ressources, sont aussi des conditions de type ($\beta$). Si le temps est discret, alors les contraintes (3.1) et (3.2) sont respectivement de la forme:

$$\sum_{i=1}^{N} v_i(t_k) \leqslant v(t_k), \tag{3.3}$$

$$\sum_{k=1}^{K} v^j(t_k) \leqslant V^j, \tag{3.4}$$

où $t_k$ est l'intervalle temporel d'indice $k$.

Traitons d'abord le cas où les conditions (β) se ramènent à des inégalités de la forme (3.1) ou (3.3).

Si les vecteurs $v(t)$ sont donnés, le plan de réalisation du projet se ramène au problème suivant: indiquer pour toute date $t$ les parts de travaux $u^i(t)$ qui doivent être exécutées pour que la date finale de réalisation de l'ensemble du projet soit la plus faible. Une telle répartition des ressources et un tel calendrier seront dits optimaux. Supposons que les parts de travaux $u^i(t)$ sont définies dans une échelle décimale et que le temps est discret. Le calendrier peut alors comme indiqué au § 2 chap. I être établi par un simple tri, puisque le nombre de variantes est fini. Mais si le nombre de travaux est

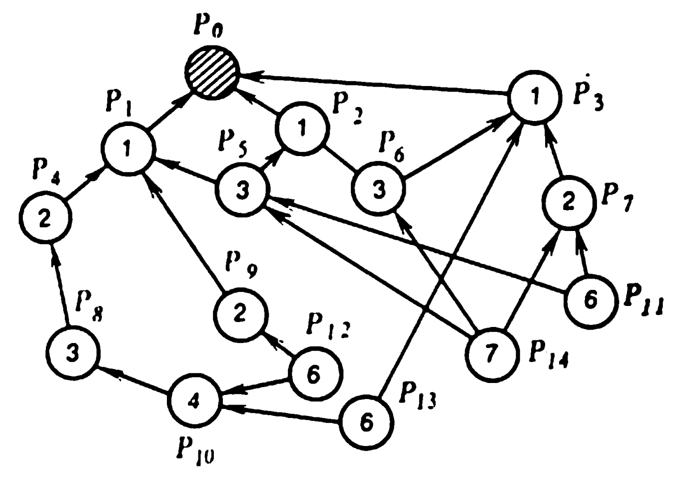



Fig. 7.2

assez élevé, alors le tri est pratiquement impossible en raison de la très longue occupation de l'ordinateur. Par ailleurs il faut que les contraintes (α) soient réalisées. Donc, nous sommes contraints pour résoudre de tels problèmes de nous rabattre sur les méthodes euristiques.

De nombreuses approches de résolution approchée des problèmes d'emploi du temps ont été proposées. Les méthodes de description agrégée sont parmi les plus importantes.

Une autre voie consiste à ordonnancer ces travaux. Si l'on réussit à ordonnancer les travaux, c'est-à-dire à leur affecter un poids et à définir un ordre d'exécution d'après leurs poids, alors on simplifie énormément la composition du calendrier. En 1960 l'auteur de cet ouvrage a proposé une variante de cette procédure euristique, appelée *ordonnancement logique*. Cette procédure affecte un grand poids à un travail antérieur à d'autres, c'est-à-dire à un travail qui conditionne le démarrage d'un grand nombre de travaux.

Expliquons le contenu de cet ordonnancement sur l'exemple que nous avons traité au chapitre I en étudiant les contraintes (α), c'est-à-dire dans le cas où les liaisons entre les travaux sont décrites par

le graphe de la figure 7.2. Supposons que les opérations $P_1$, $P_2$ et $P_3$ achèvent le projet et qu'elles sont de la même importance, donc elles sont affectées d'un même poids qui sera, par exemple, égal à l'unité puisqu'elles précèdent une seule « opération »: la fin du projet $P_0$. Les opérations $P_4$ et $P_9$ seront affectées du poids 2, car elles précèdent deux opérations: $P_1$ et $P_0$. Le même poids sera attribué manifestement à l'opération $P_7$ qui précède les deux travaux $P_3$ et $P_0$. Le poids de l'opération $P_5$ est égal à 3, car celle-ci précède les opérations $P_4$, $P_1$ et, bien sûr, $P_0$. L'opération $P_{10}$ est affectée du poids 4, car précédant les opérations $P_8$, $P_4$, $P_1$, $P_0$; l'opération $P_{11}$, du poids 6, car précédant les opérations $P_7$, $P_3$, $P_5$, $P_1$, $P_2$ et $P_0$;

*Table*

**Attribution des poids aux travaux faisant l'objet du graphe de la fig. 7.2**

| | $P_0$ | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ | $P_6$ | $P_7$ | $P_8$ | $P_9$ | $P_{10}$ | $P_{11}$ | $P_{12}$ | $P_{13}$ | $P_{14}$ | Σ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $P_0$ | | | | | | | | | | | | | | | | 0 |
| $P_1$ | 1 | | | | | | | | | | | | | | | 1 |
| $P_2$ | 1 | | | | | | | | | | | | | | | 1 |
| $P_3$ | 1 | | | | | | | | | | | | | | | 1 |
| $P_4$ | 1 | 1 | | | | | | | | | | | | | | 2 |
| $P_5$ | 1 | 1 | 1 | | | | | | | | | | | | | 3 |
| $P_6$ | 1 | | 1 | 1 | | | | | | | | | | | | 3 |
| $P_7$ | 1 | | | 1 | | | | | | | | | | | | 2 |
| $P_8$ | 1 | 1 | | | 1 | | | | | | | | | | | 3 |
| $P_9$ | 1 | 1 | | | | | | | | | | | | | | 2 |
| $P_{10}$ | 1 | 1 | | | 1 | | | | 1 | | | | | | | 4 |
| $P_{11}$ | 1 | 1 | 1 | 1 | | 1 | | 1 | | | | | | | | 6 |
| $P_{12}$ | 1 | 1 | | | 1 | | | | 1 | 1 | 1 | | | | | 6 |
| $P_{13}$ | 1 | 1 | | 1 | 1 | | | | 1 | | 1 | | | | | 6 |
| $P_{14}$ | 1 | 1 | 1 | 1 | | 1 | 1 | 1 | | | | | | | | 7 |

*idem* pour l'opération $P_{12}$, car précédant $P_{10}$, $P_9$, $P_8$, $P_4$, $P_1$ et $P_0$; *idem* aussi pour l'opération $P_{13}$ qui précède les opérations $P_{10}$, $P_8$, $P_4$, $P_1$, $P_3$ et $P_0$. Le poids de l'opération $P_{14}$ est égal à 7.

Le calcul des poids est conduit à l'aide de la table ci-dessus dont la signification évidente n'appelle aucun commentaire. A noter que ce calcul est manuel et demande peu de temps. Pour plus de suggestion il faut dresser le tableau en même temps que le graphe de la figure 7.2.

On remarque que les opérations se répartissent en groupes de même poids. Le groupe le plus « lourd » est celui qui est composé d'une seule opération $P_{14}$. Le poids de ce groupe d'opérations est égal à 7. Viennent ensuite les groupes d'opérations $P_{11}$, $P_{12}$ et $P_{13}$; $P_{10}$; $P_5$, $P_6$ et $P_8$; $P_4$, $P_9$ et $P_7$, et enfin le groupe le plus « léger » $P_1$, $P_2$ et $P_3$. Les poids respectifs de ces groupes sont 6, 4, 3, 2 et 1. Ces opérations doivent être effectuées en commençant par les plus « lourdes » (en l'occurrence $P_{14}$), celles qui freinent le démarrage de la plus grande partie des travaux.

Cet ordonnancement des opérations présente beaucoup d'avantage, mais son principal atout est que les opérations d'un même groupe sont indépendantes: on peut les accomplir dans un ordre quelconque et même simultanément. L'autre qualité importante est la possibilité de regrouper les opérations de même poids, ceci est la conséquence de leur indépendance. On obtient en définitive un graphe linéaire (fig. 7.3).

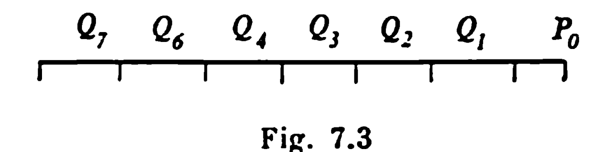



Fig. 7.3

Dans l'exemple considéré, ce sera le graphe avec les travaux $P_0$, $Q_1$, $Q_2$, $Q_3$, $Q_4$, $Q_6$ et $Q_7$. Le travail $Q_i$ est la réunion de tous les travaux de poids $i$. Les calculs nécessités par la réalisation d'un graphe d'une structure aussi simple sont élémentaires.

Donc, la marche à suivre pour dresser un calendrier de travaux est la suivante:

1. On calcule le poids des divers travaux et on remplace le graphe primitif par un graphe linéaire.
2. On mobilise toutes les ressources pour le démarrage des plus « lourds » travaux ($Q_7$ dans l'exemple envisagé).
3. Une fois les travaux $Q_7$ achevés, on branche les ressources sur les travaux $Q_6$ et ainsi de suite.
4. A l'intérieur de chaque groupe on résout des problèmes de commande optimale.

REMARQUE. Cette méthode peut être facilement améliorée au cours de sa réalisation. Si, par exemple, on accomplit les travaux $Q_n$ sans mobiliser toutes es ressources, on peut dans le cadre d'une nouvelle répartition des ressources

disponibles avancer certains travaux d'un autre groupe : on peut, par exemple, réunir les travaux du groupe $Q_n$ avec ceux du groupe $Q_{n-1}$ qui ne sont précédés d'aucun travail $Q_n$ et qui peuvent être accomplis en même temps que les travaux $Q_n$. Dans l'exemple envisagé, les travaux du groupe $Q_7$ ($P_{14}$) peuvent être effectués en même temps que ceux du groupe $Q_6$ ($P_{11}$, $P_{12}$, $P_{13}$).

Cette méthode pourrait être appelée méthode d'élimination des contraintes logiques. C'est une des nombreuses méthodes de simplification du problème. Il est évident que de telles méthodes euristiques ne permettent pas dans le cas général d'obtenir une solution exacte qui soit effectivement optimale. Mais elles fournissent toujours une majoration de la solution, ce qui est d'une importance capitale. En effet, supposons que cette méthode nous donne la date optimale $\hat{T}$ d'achèvement des travaux. La date $\hat{T}$ est toujours supérieure à la date optimale exacte $T^*$, c'est-à-dire que $T^* \leqslant \hat{T}$. Supposons par ailleurs que l'analyste commence à chercher de nouvelles méthodes euristiques qui à son avis correspondraient mieux à la nature du problème que la méthode décrite ci-dessus. Supposons qu'il réussisse à composer un nouveau calendrier à l'aide de l'autre méthode. Désignons par $T^{**}$ la date d'achèvement des travaux fixée par le nouveau calendrier. Etant donné qu'il est impossible en principe de procéder à une optimisation exacte, on aura $T^* \leqslant T^{**}$. Si $T^{**} < \hat{T}$, la nouvelle méthode est meilleure que celle développée ci-dessus.

S'agissant des problèmes d'emploi du temps, signalons encore un fait. Les problèmes de planning (où il faut déterminer les délais d'achèvement d'un projet ou son coût) n'impliquent pas en général une précision particulière pour plusieurs raisons : données initiales inexactes, ratios imprécis, perturbations inévitables, etc. De plus, la répartition des ressources adoptée peut toujours servir de première approximation pour un calcul plus exact, les améliorations ultérieures pouvant être effectuées par des méthodes moins laborieuses de la théorie des perturbations.

b) *Problèmes variationnels sur les graphes.* Voyons maintenant des problèmes un peu plus compliqués. Soient $N$ travaux (opérations) reliés par des conditions logiques ($\alpha$). L'exécution de chaque opération est décrite par une équation de la forme

$$\dot{x}_i = f_i(x_i, v_i), \quad i = 1, 2, \ldots, N, \tag{3.5}$$

où $x_i(t)$ est un vecteur de phase caractérisant l'état de l'opération, $v_i$ les ressources nécessaires à la réalisation de cette opération. Imposons aux commandes $v_i$ des conditions ($\beta$), c'est-à-dire des conditions de la forme (3.1) ou (3.2), ainsi que les contraintes

$$v_i \in G_i, \tag{3.6}$$

où $G_i$ sont des ensembles. La signification des contraintes (3.6) est assez évidente : si les travaux ne sont pas préparés, les équipements ne peuvent pas être utilisés à plein.

Pour variables de phase on peut prendre les quantités les plus diverses: la quantité de béton coulée dans le corps d'un barrage, la quantité de marchandises transportées, etc. Mais il est commode d'introduire des quantités relatives: par exemple, traiter $x_i$ (s'il est scalaire) comme la part de travail accompli. Si $x_i$ est un vecteur, on prendra ses composantes $x_i^j$ pour parts. Avec une telle interprétation des variables de phase, à la fin des travaux on aura

$$x_i^j (T) = 1. \tag{3.7}$$

La principale difficulté se trouve encore dans la nécessité de satisfaire des conditions ($\alpha$). Notre expérience de résolution des problèmes de commande optimale nous suggère un procédé naturel pour tourner cette difficulté: la méthode des fonctions de pénalisation. Traitons le cas où $x_i$ sont des fonctions scalaires. Les conditions ($\alpha$) s'écrivent

$$x_i (t) = 0 \text{ si } x_j (t) < 1. \tag{3.8}$$

Pour la condition (3.8) on peut construire une fonction de pénalisation, par exemple de la forme suivante:

$$\Phi_{ij} = \lambda_{ij} x_i^2 (1 - x_j)^2, \quad \lambda_{ij} \geqslant 0. \tag{3.9}$$

En effet, la fonction (3.9) est positive, puisque tous les $x_i \geqslant 0$ de par la position du problème et $x_j \leqslant 1$ en vertu de la condition (3.7). Elle est nulle si $x_i = 0$ ou $x_j = 1$. Donc, elle est différente de zéro si et seulement si $x_i \neq 0$ et $x_j < 1$ simultanément, c'est-à-dire si la condition (3.8) n'est pas remplie. La fonction (3.9) est dérivable, donc cela simplifiera les procédures de résolution des problèmes d'optimisation.

Si l'on se sert de fonctions de pénalisation (3.9), il faut remplacer la fonctionnelle $J(v_i)$ par une fonctionnelle de la forme

$$I = J + \int_0^T \sum_{i,\, j} \lambda_{ij} x_i^2 (1 - x_j)^2 \, dt, \tag{3.10}$$

où la somme est étendue à tous les indices $i$, $j$ pour lesquels ont lieu les contraintes (3.8).

REMARQUE. En fait, on peut réduire considérablement le nombre de fonctions de pénalisation. Il suffit pour cela de ne considérer que des travaux « jointifs ». Si, par exemple, on a affaire au graphe de la figure 7.2, il faut introduire dans la fonctionnelle les fonctions $\Phi_{12}$ et $\Phi_{23}$, mais pas obligatoirement la fonction $\Phi_{13}$, puisque l'introduction de $\Phi_{23}$ garantit que l'opération $P_2$ ne démarrera pas tant que l'opération $P_3$ ne sera pas achevée, et la présence de la fonction $\Phi_{12}$, que $P_1$ ne commencera pas tant que ne sera pas finie $P_2$, et par conséquent $P_3$.

L'introduction des fonctions de pénalisation ramène le problème à conditions logiques à un problème classique de commande optimale. En tout cas formellement. Mais ce problème reste extrêmement compliqué.

Les conditions (α) traduisent la nécessité de tenir compte de la structure du projet, qui est toujours discrète. Tous les gros projets se caractérisent par des contraintes structurelles discrètes et des contraintes dynamiques continues.

L'élimination des contraintes structurelles par les fonctions de pénalisation est un important instrument de l'analyse des systèmes. Si les variables de phase sont de dimension élevée, la résolution directe du problème de commande optimale et la composition, sur la base de cette solution, d'un calendrier optimal (d'un plan optimal) semblent relever de l'utopie. En réalité, la détermination du plan optimal doit s'appuyer sur un schéma itératif dont le premier pas doit être relativement simple. Pour construire la première approximation, on peut se servir de la procédure d'élimination des contraintes logiques développée au numéro précédent. Voyons à quoi mène cette procédure dans le cas d'un problème d'optimisation de la fonctionnelle $J$ sous les contraintes dynamiques (3.5).

Discutons tout d'abord la structure des fonctionnelles qui valent la peine d'être considérées dans de tels problèmes. Commençons par le délai de réalisation $T$ d'un projet. Les problèmes de rapidité sont tout indiqués ici : il faut répartir les ressources pour achever un projet dans les plus brefs délais, c'est-à-dire satisfaire les conditions (3.7). Mais les problèmes de rapidité ne sont pas les seuls problèmes variationnels qui se posent ici. Dans les gros projets, comme la construction de la voie ferrée BAM ou d'un important système énergétique, les délais d'achèvement des travaux se mesurent en années. Ici le problème doit être posé d'une autre manière : il faut qu'à une date $t^*$ donnée l'état des travaux soit avancé au maximum. Mais comment formaliser ce problème ? Il se trouve que l'interprétation ci-dessus des variables de phase est très commode. Les quantités $x_i^j$ étant toutes sans dimension, pour fonctionnelle $J$ on peut prendre leur somme :

$$J = \sum_{i,\, j} x_i^j (t^*). \tag{3.11}$$

On est ainsi conduit au problème suivant : trouver une répartition des ressources vérifiant des contraintes (β) et maximisant la fonctionnelle (3.11) à une date donnée $t^*$ sous réserve que les variables de phase satisfont les conditions dynamiques (3.5) et des contraintes (α). Dans ce problème aucune condition n'est imposée à l'extrémité droite de la trajectoire de phase. On pourrait envisager d'autres problèmes. Par exemple, les travaux $x_i^j$ peuvent être affectés de poids ou bien certains travaux doivent être obligatoirement achevés.

L'élimination des contraintes structurelles par un ordonnancement des travaux conduit, on l'a vu, à une partition de la séquence de travaux en groupes simultanément réalisables. Supposons qu'une telle partition a déjà été faite. On a alors les groupes suivants.

Le groupe $Q_s$ des plus « lourds » travaux que nous désignerons par $x_{1s}, x_{2s}, \ldots$

Le groupe $Q_{s-1}$ de travaux de poids $x_{1,\,s-1}, x_{2,\,s-1}, \ldots$

Le groupe $Q_{s-2}$, etc.

Etant donné que tous les travaux de chaque groupe $Q_i$ doivent être achevés à la même date, on décompose le problème variationnel initial en problèmes de commande optimale pour les travaux de chaque groupe $Q_i$. Nous devons donc trouver une répartition optimale des ressources pour réaliser les travaux de chaque groupe $Q_i$.

Que signifie dans ce cas la notion de répartition optimale des ressources? Si l'on envisage un problème de rapidité, c'est-à-dire la composition d'un calendrier qui permette de réaliser l'ensemble des travaux en un temps minimum, il est alors évident que sa décomposition donnera lieu à plusieurs problèmes de rapidité. En effet, étant donné que tous les travaux doivent finir à la même date, leur délai total $T$ de réalisation sera égal à la somme des dates $T_i$ de réalisation des travaux $Q_i$, soit

$$T = \sum_{i=1}^{s} T_i. \tag{3.12}$$

Donc, les conditions nécessaires et suffisantes d'optimalité du problème de rapidité obtenu par ordonnancement des travaux et passage à des contraintes de type graphe linéaire sont

$$T_i \Rightarrow \min, \quad i = 1, \ldots, s. \tag{3.13}$$

Nous sommes ainsi conduits aux $s$ problèmes suivants: déterminer le minimum de la fonctionnelle (3.13) sous les conditions

$$\frac{dx_i^j}{dt} = f_i^j(x_i^j, v_i^j), \tag{3.14}$$

$$x_i^j(0) = (0), \quad x_i^j(T_i) = 1, \tag{3.15}$$

$$\sum_j v_i^j \leqslant V_i(t). \tag{3.16}$$

Ces problèmes sont bien plus simples que le problème primitif. D'abord, leur dimension est environ $s$ fois plus petite. Mais ce n'est pas la plus importante simplification. Ce qui est primordial, c'est que les problèmes (3.13) à (3.16) ne contiennent plus de contraintes logiques ($\alpha$) et peuvent être traités par les méthodes classiques.

Le schéma de décomposition proposé peut être appliqué à la résolution de problèmes de commande avec une fonctionnelle diffé-

rente du délai $T$ de réalisation de l'ensemble des travaux. Soit donné, par exemple, un problème de planification pour une période donnée $\tilde{T}$ dans lequel on cherche une répartition des ressources qui maximise la fonctionnelle

$$J=\sum_i x_i(\tilde{T}). \tag{3.17}$$

Décomposons les travaux en groupes $Q_s$, $Q_{s-1}$, . . . On rappelle que les travaux du groupe $Q_i$ doivent être accomplis simultanément et que les travaux du groupe $Q_{i-1}$ ne peuvent être entamés tant que ceux du groupe $Q_i$ ne sont pas achevés.

Si $\tilde{T} > T_s$, où $T_s$ est le délai optimal de réalisation des travaux $Q_s$, on retient la répartition obtenue en résolvant le problème de rapidité (3.13) à (3.16) pour le groupe $Q_s$ et on passe à l'analyse du groupe suivant $Q_{s-1}$.

Si $\tilde{T} > T_s + T_{s-1}$, on conserve pour la réalisation des travaux $Q_{s-1}$ la répartition obtenue par la résolution du problème de rapidité pour ce groupe.

En poursuivant ce raisonnement, on aboutira à l'inégalité (si seulement le problème n'est pas dégénéré et si $\tilde{T} \leqslant T^*$, où $T^*$ est le délai optimum d'achèvement de l'ensemble des travaux)

$$\tilde{T} < T_s + T_{s-1} + \ldots + T_{s-k}. \tag{3.18}$$

Si l'inégalité (3.18) est réalisée, on procède comme suit. Pour les travaux $Q_s$, $Q_{s-1}$, . . ., $Q_{s-k+1}$, on conserve la répartition des ressources donnée par la résolution du problème de rapidité respectif. Pour les travaux $Q_{s-k}$, on résoudra le problème à horizon fixe et à extrémité droite libre suivant:

$$\left.\begin{gathered}\frac{dx^j_{s-k}}{dt}=f^j_{s-k}(x^j_{s-k}, v^j_{s-k}), \quad x^j_{s-k}(0)=(0),\\ \sum_{j\in Q_{s-k}} v^j_{s-k}(t) \leqslant V_{s-k}(t),\end{gathered}\right\} \tag{3.19}$$

$$J=\sum_{j\in Q_{s-k}} x^j_{s-k}(\tilde{T}-T_s-T_{s-1}-\ldots-T_{s-k+1}) \Rightarrow \max.$$

Ainsi, le schéma d'ordonnancement proposé permet de remplacer un problème difficile de commande optimale avec contraintes de type graphe par des problèmes bien plus simples.

c) *Eventuels procédés d'amélioration de la solution approchée.* Les procédures proposées définissent toujours une solution admissible qui n'est qu'une estimation de la solution exacte: une majoration de l'horizon dans le cas d'un problème de rapidité; une minoration, dans le cas de la fonctionnelle (3.17).

Voyons maintenant si la solution acquise peut être améliorée par une analyse du système (3.5) et par la méthode des fonctions de pénalisation.

Supposons qu'on a résolu le problème de rapidité, c'est-à-dire qu'on a réparti les ressources et composé un calendrier à l'aide de la méthode d'ordonnancement logique. On connaît donc les commandes $v_i(t)$ et le délai $\hat{T}$ d'achèvement de l'ensemble des travaux. Désignons par $T^*$ le délai optimal: $T^* \leqslant \hat{T}$.

Donnons-nous maintenant une date $\tilde{T} < T^*$. Le choix de cette date peut être délicat, car il est souhaitable que l'écart entre $\tilde{T}$ et $T^*$ ne soit pas très élevé. Nous glisserons sur la question du choix de $\tilde{T}$ et admettrons qu'un $\tilde{T} < T^*$ nous est donné. Par hypothèse, les conditions (3.7) doivent être réalisées à la date finale. Comme $\tilde{T} < T^*$, il n'existe pas de commande qui transfère le système de l'état initial à l'état (3.7) durant l'intervalle de temps $t = \tilde{T}$. On introduira donc une fonctionnelle $J$ caractérisant la distance à l'objectif de la commande. La fonctionnelle $J$ peut prendre des formes diverses; par exemple, on peut envisager la minimisation de

$$J = \sum_i (1 - x_i(\tilde{T}))^2. \tag{3.20}$$

REMARQUE. On pourrait aussi considérer la maximisation de la fonctionnelle

$$J = \sum_i x_i(\tilde{T}). \tag{3.21}$$

Pour éliminer les contraintes logiques, on se sert des fonctions de pénalisation (3.9); pour résoudre le problème auxiliaire, on se servira de la fonctionnelle (3.10) que, compte tenu de (3.20), on mettra sous la forme

$$I = -2 \int_0^{\tilde{T}} \sum_i (1 - x_i(t)) f_i(x_i, v_i)\, dt + \lambda \int_0^{\tilde{T}} \sum_{i,j} x_i^2 (1 - x_j)^2\, dt, \tag{3.22}$$

où $\lambda \gg 1$. Les fonctions de pénalisation (3.9) non seulement permettent de rejeter les contraintes ($\alpha$), mais assument une autre tâche. Elles préservent en effet les conditions

$$x_i(t) \leqslant 1 \tag{3.23}$$

qui sont vérifiées par toutes les coordonnées $x_i$.

Nous sommes ainsi conduits au problème de minimisation de la fonctionnelle (3.22) durant l'intervalle de temps $t = \tilde{T}$ sous les con-

ditions

$$\dot{x}_i = f_i(x_i, v_i),$$
$$v_i \geqslant 0, \quad \sum v_i(t) \leqslant V, \tag{3.24}$$
$$x_i(0) = x_0.$$

Le problème (3.22), (3.24) est à extrémité droite libre. Nous admettrons que le problème de commande optimale relatif à l'ordonnancement des travaux a déjà été résolu, donc que nous connaissons ses solutions $x_i^0(t)$ et $v_i^0(t)$. Nous disposons ainsi d'une bonne première approximation. Ceci nous suggère d'utiliser une méthode classique pour résoudre le problème (3.22), (3.24), par exemple la méthode d'approximations successives de Krylov-Tchernoousko développée au chapitre II. Reprenons ces raisonnements pour l'exemple envisagé. Composons le hamiltonien

$$H = \sum_i \psi_i f_i(x_i, v_i) + 2\sum_i (1 - x_i) f_i(x_i, v_i) - \lambda \sum_{i, j} x_i^2 (1 - x_j)^2. \tag{3.25}$$

On rappelle que la dernière somme de (3.25) est étendue à tous les indices $i$ et $j$ tels que $i$ précède immédiatement $j$. Les impulsions $\psi_i$ doivent vérifier les équations

$$\dot{\psi}_i = -\frac{\partial H}{\partial x_i} = -\psi_i \frac{\partial f_i}{\partial x_i} + 2 f_i(x_i, v_i) -$$
$$- 2(1 - x_i)\frac{\partial f_i}{\partial x_i} + 2\lambda \sum{}' x_i (1 - x_j)^2 - 2\lambda \sum{}'' x_k^2 (1 - x_i), \tag{3.26}$$

où $\sum'$ est étendue aux indices $j$ suivant immédiatement $i$, la somme $\sum''$, aux indices $k$ précédant immédiatement $i$. Vu que nous envisageons un problème à une extrémité libre, les impulsions $\psi_i(\widetilde{T})$ satisfont des conditions de transversalité nulles

$$\psi_i(\widetilde{T}) = 0. \tag{3.27}$$

Le premier pas de la méthode de Krylov-Tchernoousko consiste à résoudre le problème de Cauchy suivant :

$$\dot{x}_i = f_i(x_i, v_i^0), \tag{3.28}$$
$$x_i(0) = x_0. \tag{3.28'}$$

La résolution du problème (3.28), (3.28′) nous donne la trajectoire $x_i^0(t)$ et par suite la valeur terminale $x_i^0(\widetilde{T})$. On calcule en même temps la valeur correspondante de la fonctionnelle $I^0$. A noter que, la commande $v_i^0(t)$ étant admissible, la trajectoire $x_i^0(t)$ satisfera toutes les contraintes, y compris les contraintes logiques ($\alpha$). Donc, le deuxième terme de (3.22) sera nul. Nous pouvons résoudre mainte-

nant le problème de Cauchy

$$x_i(\tilde{T}) = x_i^0(\tilde{T}), \quad \psi_i(\tilde{T}) = 0$$

relatif aux systèmes d'équations (3.26), (3.28) où l'on conviendra de poser $v_i = v_i^0$. Pour obtenir la solution $v_i^1$ de ce problème on se servira du principe du maximum. Ce problème se trouve ainsi ramené au problème de programmation non linéaire

$$H^* = \sum \psi_i f_i(x_i, v_i) + 2\sum (1 - x_i) f_i(x_i, v_i) \Rightarrow \max, \qquad (3.29)$$

où $v_i$ satisfont les contraintes $\sum v_i \leqslant V$.

Une fois qu'on a résolu le problème (3.29) et déterminé la nouvelle commande $v_i^1$, on peut reprendre la procédure décrite. On peut en particulier calculer la nouvelle valeur de la fonctionnelle $I^1$.

Si $I^1 < I^0$, on obtient une meilleure valeur de la commande. Mais le problème n'est pas linéaire et dans le cas général il est possible que $I^1 \geqslant I^0$. Il faut alors procéder autrement. Introduisons la quantité $\Delta v^0(m) = (v^1 - v^0)/m$ et posons $\tilde{v}^1 = v^0 + \Delta v^0(m)$ et choisissons ensuite $m$ tel que $\tilde{I}^1 < I^0$.

Si $\tilde{T}$ et $\hat{T}$, qui sont reliées au temps optimal $T^*$ par la double inégalité

$$\tilde{T} < T^* \leqslant \hat{T},$$

sont assez voisines du point de vue de l'expert (par exemple de l'analyste du système), on peut mettre un terme aux calculs : on prend la commande $v^1$ entre 0 et $\tilde{T}$ et la commande $v^0$ pour $t > \tilde{T}$. Si la date $\tilde{T}$ n'est pas assez proche de $\hat{T}$, c'est-à-dire si $\hat{T} - T^*$ est grand, nous devons recommencer la procédure décrite pour $T_1 > \tilde{T}$, mais $T_1 < T^*$.

d) *La méthode de programmation et problèmes de graphes.* Les problèmes considérés dans ce paragraphe sont d'une importance capitale pour les applications et notamment pour la réalisation de la méthode de programmation.

En effet, un problème central de la méthode de programmation est la composition des programmes, leur coordination, une répartition raisonnable des ressources et la commande de ces programmes. Il est d'usage de distinguer les programmes de développement de grande envergure : construction de nouvelles entreprises, d'unités industrielles, extension des transports, du potentiel défensif, de l'enseignement, etc. En général, ces programmes sont des listes de travaux reliés par diverses conditions logiques et en premier lieu par des conditions de type graphe. A chaque travail il faut associer des ressources : si l'on utilise la terminologie du paragraphe précédent, il faut se donner des fonctions $f_i(x_i, v_i)$. On convient de dire

qu'un programme est donné (formulé) si sont définis les ressources et les liens logiques entre les travaux.

Dans ce paragraphe on a admis que les contraintes de type ressources étaient exogènes. Mais les ressources sont créées par le travail des entreprises. En d'autres termes, il existe deux processus dynamiques: la réalisation du programme et la création des ressources nécessaires à sa réalisation. Ces deux processus sont étroitement liés entre eux. Nous avons donc envisagé un seul aspect de ce phénomène complexe.

Certes, essayer de formuler, et *a fortiori* d'attaquer de front, le problème de réalisation optimale d'un programme relève de l'utopie. Il faut faire appel ici à des procédures itératives. Les problèmes d'optimisation décrits au n° b) peuvent servir d'éléments de telles procédures itératives.

Il existe encore un problème qui occupe une place importante dans les procédures de la méthode de programmation : le problème de la commande du programme. Pour composer le calendrier des travaux, on est parti des équations déterministes (3.5). En réalité, les paramètres du processus ne sont tous assez exactement connus et les perturbations extérieures, rigoureusement déterminées. Bref, le système d'équations (3.5) devrait être remplacé par le système

$$\dot{x}_i = f_i(x_i, v_i, \xi_i), \tag{3.30}$$

où $\xi_i$ est un facteur indéterminé ou aléatoire. Etant donné que la commande programmée sur la base de laquelle nous avons dressé le calendrier a été acquise grâce à des équations déterministes, nous devons envisager une rétroaction corrective pour réaliser le programme.

Le problème de commande de travaux liés par des conditions de type graphe orienté est à certains égards classique : il a été posé dès les années cinquante et il a fait l'objet d'une foule de travaux. Dans le cadre de ce problème sont apparues de nombreuses méthodes intéressantes d'organisation de la rétroaction corrective, la plus importante étant la *méthode du chemin critique* dite encore méthode P.E.R.T.

Supposons que nous avons dressé un calendrier des travaux. A chaque travail $P_i$ est alors associé son délai d'achèvement $T_i$. Supposons maintenant que l'on possède un instantané de l'état des travaux réalisés, c'est-à-dire qu'on connaît les quantités $x_i(t^*)$, où $t^*$ est une date fixe. En général les quantités $x_i(t^*)$ peuvent être différentes de leurs valeurs programmées $x_i^*(t^*)$. Cet écart est dû précisément à l'action des facteurs aléatoires négligés.

L'« état des travaux » $x_i(t^*)$ définit un ensemble de points $\mathcal{P}_i$ sur le graphe. Chacun de ces points est relié au sommet du graphe (au travail $P_0$) par un nombre fini de chemins représentant une partie des travaux. La date $T_i$ d'achèvement de chaque travail étant connue, on peut calculer le délai total de réalisation des travaux de chaque

chemin, c'est-à-dire des séquences de travaux reliant les points $\mathcal{P}_t$ au sommet $P_0$. Parmi ces séquences il en existe au moins une dont le temps de réalisation, d'après le calendrier initial, sera maximal. Ces séquences sont appelées *chemins critiques*. Ce sont eux précisément qui définiront la date finale d'achèvement des travaux et reculeront la fin du projet. Si l'on réussit à réduire les délais de réalisation des travaux des chemins critiques, on réduira *ipso facto* la date d'achèvement de l'ensemble des travaux. Cette propriété du chemin critique a servi de base à la création du premier système de gestion. Le dispatcher suit non pas l'ensemble de tous les travaux, mais seulement ceux qui se trouvent sur les chemins critiques. Grâce à l'information qu'il reçoit sur l'état des travaux, il calcule rapidement les chemins critiques, les retards temporels, retards qu'il essaye de combler en activant les secteurs traînants.

Il existe actuellement plusieurs méthodes de calcul du chemin critique et les programmes correspondants sont inclus dans le logiciel de la plupart des ordinateurs.

La méthode du chemin critique n'est pas la seule méthode d'organisation de la rétroaction. Au chapitre II on a exhibé une méthode de synthèse appelée méthode du plan glissant qui utilise le même algorithme de calcul que le programme optimal. Etant donné que l'ordonnancement logique des travaux et l'amélioration des commandes nous fournissent des méthodes assez économiques de calcul des programmes, on peut les appliquer à la réalisation du programme. Les étapes de cette correction sont les suivantes:

1. On calcule le programme optimal: on trouve les commandes $v_i^*(t)$ et la trajectoire de phase $x_i^*(t)$.

2. Au bout d'un intervalle de temps $\tau$ qui est un paramètre du système (et qui est défini par des facteurs techniques) on mesure les variables de phase $x_i(t)$. Etant donné que le calcul a été effectué à l'aide des équations (3.5) et que le processus réel est gouverné par les équations (3.30), on obtient le désaccord

$$\delta_i(\tau) = x_i(\tau) - x_i^*(\tau). \tag{3.31}$$

Dans cette étape, on précise aussi les éventuelles indéterminations, par exemple les ratios. En d'autres termes, on remplace le système (3.5) par le système (3.30) qui est plus conforme à la réalité. En résolvant les équations (3.30) en tenant compte du nouvel état initial $x_i(\tau) = x_i^*(\tau) + \delta_i(\tau)$, on obtient un nouveau programme et l'on définit de nouvelles commandes.

3. Les travaux sont accomplis d'après le nouveau calendrier entre les dates $\tau$ et $2\tau$, et ainsi de suite.

Ce schéma exige, en comparaison du schéma de calcul de la commande optimale, un ordinateur plus puissant, mais il fournit à l'analyste une information bien plus complète et permet par conséquent d'améliorer nettement la commande.

## § 4. Problèmes de synthèse matricielle

a) *Position et interprétations du problème.* On traite dans ce paragraphe un système gouverné par des équations à temps discret de la forme

$$x(t_{n+1}) = f_n(x(t_n),\ u(t_n),\ \xi(t_n)), \qquad n = 0,\ 1,\ \ldots,\ N, \tag{4.1}$$

où $u(t_n)$ est la commande, $\xi(t_n)$, des facteurs aléatoires. Les commandes sont choisies de manière à minimiser la fonctionnelle additive

$$J = \sum_{n=0}^{N-1} F_n(x_n,\ u_n) + F_N(x_N). \tag{4.2}$$

L'état initial du système

$$x(t_0) = x_0 \tag{4.3}$$

est généralement une quantité aléatoire. La fonction $x(t_n)$ est une fonction vectorielle aléatoire, puisque les seconds membres des équations (4.1) contiennent des fonctions aléatoires de l'argument discret $\xi(t_n)$ et l'état initial $x_0$ est aléatoire aussi. Donc, la fonctionnelle (4.2) est aléatoire. On admettra dans la suite qu'il faut minimiser l'espérance mathématique de cette fonctionnelle, soit

$$I = \overline{J}. \tag{4.4}$$

Les problèmes (4.1) à (4.4) trouvent un très vaste champ d'application en analyse des systèmes. En analysant les équations (4.1) il faut nécessairement prendre en ligne de compte et utiliser les propriétés physiques des systèmes envisagés. Dans ce paragraphe on se propose d'étudier la commande d'une cascade de barrages à l'aide d'un système de la forme (4.1).

Désignons par $x^i(t_n)$ la capacité du barrage $i$ à l'instant $t_n$; par $u^i(t_n)$, la lâchure à la date $t_n$ du barrage $i$ dans le barrage $i+1$; par $R^i(t_n)$, l'apport secondaire, c'est-à-dire la quantité d'eau reçue par le cours entre les barrages $i$ et $i+1$ à la date $t_n$; par $S^i(t_n)$, la quantité d'eau prise du barrage $i$ à la même date pour des besoins économiques et notamment pour l'irrigation. Les quantités $u^i(t_n)$ et $S^i(t_n)$ sont commandables. L'apport $R^i(t_n)$ est une quantité aléatoire incontrôlable. La relation d'équilibre fondamentale régissant le fonctionnement de ce système de barrages est de la forme

$$x^i(t_{n+1}) = x^i(t_n) + u^{i-1}(t_n) - u^i(t_n) + R^i(t_n) - S^i(t_n). \tag{4.5}$$

L'état initial du système est

$$x^i(0) = x_0^i. \tag{4.6}$$

Les quantités $x_0^i$ sont supposées aléatoires. Le choix des commandes dépend de nombreux objectifs: quantité d'électricité produite, superficie et rendement des terres à irriguer, etc. La première étape consiste à réduire ces critères. On glissera sur ces procédures qui ont été évoquées au chapitre I. On admettra donc que la fonctionnelle a été désignée et qu'elle est de la forme

$$J=\sum_i \sum_n F_i(x^i(t_n),\ x^i(t_{n-1}),\ R^i(t_n))+\sum_i \sum_n \Phi_i(S^i(t_n),\ \eta^i(t_n)), \tag{4.7}$$

où $\eta^i(t_n)$ est une quantité aléatoire caractérisant la pluviosité. La fonctionnelle (4.7) se ramène à la forme (4.2), puisque la commande $u(t_{n-1})$ s'exprime en fonction de $x(t_{n-1})$ et $x(t_n)$ grâce à l'équation $u(t_{n-1})$.

REMARQUE. Il existe actuellement d'innombrables modèles de systèmes d'utilisation de l'eau qui comprennent aussi le système des barrages. Ces modèles se distinguent essentiellement par la description des critères d'efficacité, c'est-à-dire des fonctions $F_i$ et $\Phi_i$, puisque les relations d'équilibre sont d'une forme assez classique. Il est vrai qu'elles peuvent être légèrement modifiées par des facteurs secondaires, par exemple par le « temps de décalage », ou autres. Mais ces perfectionnements du modèle laissent intacte sa « classe d'appartenance », autrement dit le choix des commandes se ramène à la résolution de problèmes de la forme (4.1) à (4.3).

Quels problèmes de commande se posent alors? Des problèmes classiques: le choix d'un programme et la construction d'un système de commandes correctives réalisant la rétroaction. Traitons maintenant ces deux problèmes à tour de rôle.

b) *Choix d'une commande programmée.* D'après la méthode de programmation, pour résoudre le problème posé il faut d'abord se donner les valeurs des facteurs aléatoires et incontrôlables $\xi^i(t_n)$ et $x^i(0)$. Dans la gestion des barrages, ces facteurs sont nombreux: niveau de l'eau, apport secondaire, niveau des précipitations, etc. Le niveau de l'eau dans le réservoir peut être donné par le traitement des observations pluriannuelles ou par une expertise; l'apport secondaire et le niveau des précipitations, par une prévision.

Ingénieurs et analystes accordent une grande importance à la prévision comme le témoignent les innombrables travaux réalisés sur les ordinateurs les plus modernes. A mon sens, l'intérêt manifesté par les ingénieurs n'est pas toujours justifié, puisqu'on peut se contenter d'estimations grossières pour les besoins pratiques. Pour appuyer cette thèse je citerai encore trois arguments.

Premièrement, les processus aléatoires envisagés ne sont pas stationnaires et en principe leurs mesures sont entachées d'erreurs. L'application des méthodes mathématiques éprouvées au traitement de ces informations conférera *ipso facto* le sceau de la rigueur à des résultats somme toute peu sûrs.

Deuxièmement, en admettant que les $\xi^i$ sont aléatoires et en les traitant par des méthodes statistiques, nous faisons des suppositions qui traduisent la mauvaise maîtrise des causes dont dépendent les processus aléatoires envisagés. En effet, l'apport secondaire, par exemple, est la résultante de plusieurs facteurs, tels l'hydraulicité de l'année précédente, la quantité de neige tombée en hiver, la quantité d'eau puisée dans les nappes aquifères, etc. Or, nous n'avons pas de modèles satisfaisants décrivant les corrélations entre l'hydrodynamique souterraine et le ruissellement, nous n'avons pas non plus de modèles du ruissellement, nous ignorons de nombreuses autres causes de l'apport secondaire. Donc, nous devons le paramétriser, le traiter comme un processus aléatoire et sur la base de cette convention traiter l'information empirique. Cette paramétrisation est-elle justifiée? Bien sûr que non. Vaut-il alors la peine d'appliquer des méthodes de traitement aussi compliquées pour cette paramétrisation ?

Et enfin troisièmement, notre objectif, rappelons-le, est d'obtenir seulement une trajectoire d'appui, de trouver une commande optimale dans une « situation classique ». Revenons maintenant sur la « philosophie » de cet ouvrage. Dans les problèmes « bien » posés, les valeurs des fonctionnelles ne sont pas aux voisinages de leurs valeurs extrémales très sensibles aux petites variations des commandes. Cette circonstance a servi de base à la création des algorithmes rapides qui utilisaient des modèles assez grossiers pour la détermination des commandes. Dans le modèle envisagé il n'existe aucun petit paramètre explicite qui permette de bâtir une théorie asymptotique. Le modèle est simplifié par une paramétrisation grossière des variables et par une simplification de la procédure de traitement des informations. Il est évident que la procédure simplifiée ne sert qu'à déterminer une commande optimale. Une fois qu'on obtient une commande admissible, on la porte dans le système « exact » primitif (4.1) et on calcule la fonctionnelle en tenant entièrement compte de tous les détails du processus. Ce problème n'est plus difficile car il se ramène à un problème de Cauchy. Signalons que tout ce qui vient d'être dit concerne un système donné de valeurs $\xi\ (t_n)$.

L'étape qui suit le choix de la commande optimale est l'analyse de sa stabilité par rapport à l'information initiale. Une fois qu'on a retenu une commande $u\ (t_n)$ on peut faire varier $\xi\ (t_n)$ dans certaines limites et voir comment ces variations retentissent sur les valeurs de la fonctionnelle $J$. Ces expériences numériques qui ne demandent pas beaucoup d'efforts et de temps machine sont très utiles, car elles permettent à l'analyste de définir les limites d'applicabilité de la commande optimale. Ces limites ne seront pas *a priori* très vastes. Dans l'exemple envisagé, les commandes sont les quantités d'eau utilisées pour l'irrigation et la production de l'énergie électrique. Or ces quantités dépendent fortement de la pluviosité annuelle, de

l'hydraulicité, etc. Donc, la commande trouvée n'est qu'un jalon.

c) *Construction de l'opérateur de rétroaction.* La résolution du problème de commande optimale n'est pas une procédure laborieuse. Dans l'exemple envisagé, c'est un problème de commande optimale à une extrémité libre et le calcul approché de la trajectoire implique la résolution de plusieurs problèmes de Cauchy. Ceci nous suggère d'utiliser le même schéma de calcul que pour la construction de l'opérateur de rétroaction par la méthode du plan glissant.

Décrivons le processus de construction des opérateurs de prédiction. Pour construire la commande optimale on s'est servi de la prédiction des quantités $\xi(t_n)$ qui ont été identifiées à leurs moyennes. Mais c'était une procédure de choix de certaines moyennes ou, comme on dit encore dans ces cas, des valeurs les « plus probables » de l'apport secondaire et des caractéristiques météorologiques: les quantités $\bar{\xi}$. La situation a légèrement changé maintenant. Nous connaissons la valeur $\xi(t_0)$ à l'instant initial $t_0$ et en général $\xi(t_0) \neq \bar{\xi}(t_0)$. Cette information permet d'obtenir une nouvelle prédiction $\xi_0(t_n)$:

$$\xi_0(t_n) = \Gamma_0[\xi(t_0)], \qquad n = 1, \ldots, N. \tag{4.8}$$

(On peut en particulier utiliser une prévision météorologique à court terme et une expertise de l'apport secondaire.) A l'aide de l'opérateur (4.8), c'est-à-dire de la nouvelle prédiction, on détermine un nouveau programme et la commande $u(t_0)$ que l'on utilisera sur le premier intervalle de temps. On mesure le vecteur $x(t_1)$ et la quantité aléatoire $\xi(t_1)$ à l'instant $t = t_1$. On établit une nouvelle prédiction

$$\xi_1(t_n) = \Gamma_1[\xi(t_0), \xi(t_1)], \qquad n = 2, 3, \ldots, N,$$

puis on élabore un nouveau programme et ainsi de suite.

La construction des opérateurs de prédiction $\Gamma_i$ est un problème toujours difficile. Pour les raisons déjà signalées les méthodes mathématiques même les plus perfectionnées sont stériles. Pourtant on peut obtenir une prédiction assez bonne sur un ou plusieurs intervalles temporels sur la base d'une expertise. Donc, nous pouvons toujours nous référer aux propriétés locales du processus aléatoire $\xi(t_n)$ et à ses valeurs moyennes (« les plus probables »). Cette procédure est manifestement euristique. Mais elle a peu de chance d'être justifiée, car on ne connaît pas suffisamment bien les mécanismes qui conditionnent la formation du processus $\xi$. Avec un tel niveau d'information, on pourra tout au plus parler d'une « vraisemblance des raisonnements ».

La méthode du plan glissant réclame des ordinateurs très puissants. C'est pourquoi dans les systèmes d'utilisation de l'eau on se sert d'un autre schéma de construction de l'opérateur de rétroaction. Le dispatcher préposé au barrage $i$ observe l'état de ce dernier à l'instant $t_n$, c'est-à-dire la quantité $x^i(t_n)$. Il dispose d'un graphique

de lâchures $u^i(t_n)$ et de quantités d'eau $S^i(t_n)$ à déverser du barrage $i$. Ce procédé de commande est élémentaire. Il ne requiert pas de formation spéciale du personnel. Le régime des lâchures est décrit par une fonction de la forme

$$u^i(t_n) = \Psi(x^i(t_n),\ t_n); \tag{4.9}$$

il est calculé à l'avance et comporte une boucle de rétroaction. Ce procédé de commande présente toutefois un grave défaut : la commande du barrage $i$ est déterminée uniquement sur la base de l'information recueillie exclusivement sur lui. Quel que soit le degré de crédibilité de la fonction (4.9), le fait qu'elle ne tienne pas compte les corrélations entre les barrages déprécit considérablement cette approche. Essayons de traiter ce problème de la théorie de synthèse de la commande.

Supposons que le système (4.1) est linéaire et que la fonctionnelle (4.2) est quadratique. La commande corrective optimale sera (cf. chap. II) une fonction linéaire des écarts des variables de phase par rapport à leurs valeurs moyennes :

$$\delta u(t_n) = A_n \delta x(t_n), \tag{4.10}$$

où

$$\delta x(t_n) = x(t_n) - x^0(t_n), \qquad \delta u(t_n) = u(t_n) - u^0(t_n),$$

où $u^0$ et $x^0$ sont respectivement la commande et la trajectoire optimales obtenues sous réserve que $\xi = \bar{\xi}$ pour tous $t_n$ ; $A_n = (a_{ij}(t_n))$, la matrice de rétroaction dont les éléments $a_{ij}(t_n)$ sont les coefficients d'amplification. A noter que dans le cas considéré la commande $u(t_n)$ dépend de $x(t_n)$ et pas du passé du processus et des valeurs des vecteurs aléatoires $\xi(t_n)$, mais chaque composante du vecteur commande dépend de toutes les composantes du vecteur $x$. Ce résultat rigoureux a valeur de théorème *).

Mais dans le cas général le système (4.1) n'est pas linéaire et la fonctionnelle (4.2), pas quadratique. Il n'empêche qu'on cherchera la commande corrective sous la forme (4.10). Dans le cas général elle ne sera pas optimale, mais elle sera d'autant plus proche d'une commande optimale que la non-linéarité de l'équation (4.1) sera faible et la fonctionnelle (4.2), proche d'une fonctionnelle quadratique. Néanmoins ce sera une commande qui réalisera la rétroaction. La commande (4.10) avec des matrices $A_n$ dûment choisies sera une commande optimale, c'est-à-dire satisfaisant les contraintes requises.

Ainsi, la synthèse matricielle consiste à chercher la commande dans la classe des fonctions linéaires de variables définissant l'écart de l'état du système par rapport à la trajectoire programmée. Cette situation est typique pour les problèmes d'ingénieurs de la théorie de la commande.

*) Nous en avons déjà parlé au chapitre II.

27—01523

Avant de passer à la description des méthodes de réalisation numérique de la synthèse linéaire, arrêtons-nous sur les méthodes de calcul de l'efficacité du système. Supposons qu'on a trouvé une commande (4.10). Pour estimer son efficacité nous devons déterminer l'espérance mathématique de la fonctionnelle $J$. Théoriquement on doit porter cette commande dans l'équation (4.1):

$$x(t_{n+1}) = f_n(x(t_n), u^0(t_n) + A_n(x(t_n) - x^0(t_n)), \xi(t_n)), \quad (4.11)$$

et ensuite se donner à l'aide de la méthode de Monte-Carlo un système de vecteurs $\xi_k(t_n)$ et des conditions initiales $x_k(t_0)$, résoudre le problème de Cauchy le nombre de fois nécessaire, calculer pour chaque système $\xi_k(t_n)$ la quantité

$$J_k = J(x_k(t_0), \quad \xi_k(t_0), \ldots, \xi_k(t_{n-1})) = J_k(\zeta)$$

et trouver son espérance mathématique

$$\bar{J} = \frac{1}{K} \sum_{k=1}^{K} J_k, \quad (4.12)$$

où $K$ est le nombre total de réalisations, $k$ le numéro de la réalisation. Calculons $K$. Soient $s$ la dimension du vecteur $\xi$, $m$, celle du vecteur $x$. La dimension du vecteur $\zeta$ sera alors

$$[\zeta] = s \times N + m. \quad (4.13)$$

Désignons par $r$ le nombre d'essais nécessaires pour des raisons de précision à chaque composante du vecteur $\zeta$. Le nombre total $K$ d'essais qu'il faut effectuer pour réaliser la méthode de Monte-Carlo avec la précision donnée sera

$$K = r^{[\zeta]}. \quad (4.14)$$

Signalons que chaque essai donne lieu à la résolution d'un problème de Cauchy. Les formules (4.13) et (4.14) montrent que ces essais impliquent un temps machine astronomique. Le calcul de $\bar{J}$ à l'aide de la formule (4.12) est impossible même pour les problèmes les plus modestes.

Donc, pour évaluer l'efficacité de la commande il faut se tourner vers l'analyse des tests que seul un expert pourra désigner.

Enfin, la minimisation de la fonctionnelle $J$ peut être posée de la manière suivante. Supposons que la commande est de la forme (4.10) et prenons pour nouvelles commandes les quantités $x(t_0)$ et $\xi(t_n)$ en leur imposant des conditions de la forme $\xi(t_n) \in G(t_n)$. La résolution du problème

$$J(\xi) \Rightarrow \min_{\xi \in G},$$

nous permet de voir comment la commande corrective assume sa fonction dans les situations critiques.

REMARQUE. L'étude de l'efficacité a pour objectif essentiel de voir comment fonctionne le système. Elle réclame de grands efforts à l'analyste et une longue occupation de l'ordinateur, donc le programme définitif d'essais du système est une sorte de compromis entre les ressources investies dans l'étude des propriétés du système et l'appréhension de ses capacités.

d) *Résolution numérique d'un problème de synthèse matricielle.* On cherchera la commande sous la forme (4.10). Ceci nous permet de transcrire les conditions initiales du problème (4.1), (4.2) sous la forme

$$x(t_{n+1}) = f_n(x(t_n),\ u^0(t_n) + A_n(x(t_n) - x^0(t_n)),\ \xi(t_n)), \quad (4.15)$$

$$J = \sum_{n=0}^{N-1} F_n(x(t_n),\ u^0(t_n) + A_n(x(t_n) - x^0(t_n))) + F_N(x_N). \quad (4.16)$$

Le problème de synthèse s'énonce maintenant sous la forme suivante: déterminer la matrice $A_n$ de manière à maximiser la fonctionnelle $I$ ($I = \overline{J}$) sous les conditions (4.15) et la condition initiale $x(t_0) = x_0$, où $x_0$ est un vecteur aléatoire.

Les inconnues de ce problème sont les éléments des matrices $A_n$. Soient $m$ et $k$ les dimensions respectives des vecteurs $x$ et $u$. Le nombre des inconnues $a_{ij}(t_n)$ sera alors égal à $m \times k \times N = K^*$. Le nombre $K^*$ peut être très grand. Donc la détermination directe des quantités $a_{ij}$ à partir des conditions d'un problème de programmation stochastique peut s'avérer assez délicate. Donc, la procédure numérique doit être basée sur un schéma de décomposition. L'additivité de la fonctionnelle (4.16) plaide pour une modification de la méthode de programmation dynamique.

Considérons le dernier pas du processus. Supposons que le système se trouve dans l'état $x = x(t_{N-1})$ et posons le problème suivant: déterminer une commande $u(t_{N-1})$ réalisant le maximum de la fonctionnelle

$$I_N = \overline{F_N(x(t_N))}$$

sous la condition

$$x(t_N) = f_{N-1}(x(t_{N-1}),\ u(t_{N-1}),\ \xi(t_{N-1})). \quad (4.17)$$

Utilisons cette relation pour mettre la fonctionnelle $I_N$ sous la forme

$$I_N = \overline{F_N(f_{N-1}(x(t_{N-1}),\ u(t_{N-1}),\ \xi(t_{N-1})))} = \overline{F_N^*(x(t_{N-1}),\ u(t_{N-1}),\ \xi(t_{N-1}))}.$$

Posons ensuite

$$\xi(t_{N-1}) = \xi^0(t_{N-1}) + \eta(t_{N-1}),$$

où $\xi^0(t_{N-1})$ est une valeur prédite de la variable aléatoire $\xi(t_{N-1})$ qui sera identifiée à son espérance mathématique. Donc, $\eta(t_{N-1})$

est une variable aléatoire centrée. Traitons $F_N^* (x (t_{N-1}), u (t_{N-1}), \xi (t_{N-1}))$ comme une fonction de $\eta (t_{N-1})$ et approchons cette fonction par la parabole

$$F_N^* \sim b_0 + B_1\eta + (\eta, B_2\eta),$$

donc

$$I_N = b_0 + \overline{(\eta, B_2\eta)} = I_N (x (t_{N-1}), u (t_{N-1})), \qquad (4.18)$$

où

$$b_0 = F_N^* (x (t_{N-1}), u (t_{N-1}), \xi^0 (t_{N-1})),$$

$$\overline{(\eta, B_2\eta)} = \sum_{i, j} b_{ij} (x (t_{N-1}), u (t_{N-1}), \xi^0 (t_{N-1})) \overline{\eta^i (t_{N-1}) \eta^j (t_{N-1})}.$$

La situation qui se présente ici n'implique pas seulement une analyse formelle. Vu que nous nous intéressons à l'espérance mathématique de la fonction $F_N (x (t_N))$, il nous suffit de résoudre un problème de la forme

$$I_N \Rightarrow \max_{u(t_{N-1}) \in G} . \qquad (4.19)$$

Ce problème n'est pas très compliqué, puisqu'il est de même dimension $k$ que la commande $u$, et que $x (t_{N-1})$ est supposé fixe. Mais pour le résoudre il faut connaître la matrice des covariances (la matrice des moments d'ordre deux):

$$\Gamma (t_{N-1}) = \overline{(\eta^i\eta^j)}. \qquad (4.20)$$

Nous avons déjà discuté le problème de traitement des observations. Pour déterminer une commande optimale, nous aurions dû trouver l'espérance mathématique $\bar{\xi}$, ce qui en soi n'est pas une tâche de tout repos. Il est encore plus difficile de calculer la matrice des covariances. Dans notre exemple il faut établir l'interaction des apports secondaires dans les zones des barrages $i$ et $j$. Il est encore plus difficile de lier ces phénomènes aux caractéristiques prévisionnelles qui influent sur eux d'une manière assez complexe: cette relation est définie par la structure des nappes aquifères, les propriétés de l'hydrodynamique souterraine et par d'autres facteurs, bref par un système de corrélations réparties dans le temps et dans l'espace. Donc, la détermination des variances sous l'hypothèse que tous les processus $\xi (t_n)$ sont purement aléatoires peut se solder par des erreurs plus graves qu'une simple proscription des espérances mathématiques conjointes $\overline{\eta^i\eta^j}$.

L'analyste est donc confronté à l'alternative: ou bien négliger la composante de la fonctionnelle $I$ qui dépend des moments d'ordre deux, ou bien se servir des expertises des moments d'ordre deux.

Quelle que soit la décision prise, on est conduit à un problème d'optimisation des composantes du vecteur $u (t_{N-1})$ pour des valeurs

données des composantes du vecteur $x(t_{N-1})$, c'est-à-dire à un problème de construction de la fonction de synthèse

$$u(t_{N-1}) = u_{N-1}(x(t_{N-1})). \tag{4.21}$$

La fonction (4.21) se représente par un graphique généralisé. Le dispatcher du barrage $i$ fixe le niveau des lâchures sur le vu des observations de son barrage et des informations sur les états des autres.

REMARQUE. 1. Nous avons employé le terme « graphique généralisé ». En effet, il se distingue qualitativement des graphiques généralement utilisés, puisque la fonction (4.21) décrit la dépendance des commandes dans la zone du barrage considéré par rapport aux états des autres barrages.

2. Si le dispatcher dispose d'un ordinateur, son action est plus efficace. Il connaît la quantité $\xi(t_{N-1})$ et l'état $x(t_{N-1})$. Donc, tous les termes de l'équation (4.17) sont connus et le dispatcher peut se dispenser de calculer l'espérance mathématique. Il résoudra directement le problème de maximisation de la fonction $F_N(x(t_N))$. Il faudra tenir compte de cette particularité de la dernière tranche de la trajectoire de phase en élaborant les instructions d'exploitation.

Revenons maintenant à la construction du graphique (4.21). Approchons $u(t_{N-1})$ par la fonction linéaire

$$u(t_{N-1}) = u^0(t_{N-1}) + A_{N-1}(x(t_{N-1}) - x^0(t_{N-1})), \tag{4.22}$$

où $u^0$ et $x^0$ sont la commande et l'état du système correspondant à la trajectoire optimale.

Mettons l'équation (4.22) sous la forme scalaire:

$$u^i(t_{N-1}) = u^{0i}(t_{N-1}) + \sum_j a_{ij}(t_{N-1})[x^j(t_{N-1}) - x^{0j}(t_{N-1})]. \tag{4.22'}$$

Considérons la maximisation de la fonctionnelle (4.18). En portant (4.22') dans la fonctionnelle (4.18), on obtient une fonction non linéaire des coefficients d'amplification $a_{ij}$ et le problème d'optimisation est de dimension $m^2$. Mais ce problème peut être simplifié si l'on se sert d'une forme spéciale de la dépendance (4.22) de la commande par rapport au vecteur de phase. Supposons par exemple que le vecteur $x(t_{N-1})$ est défini comme suit:

$$\begin{aligned} &x^i(t_{N-1}) = x^{0i}(t_{N-1}) + \delta^i, \\ &x^j(t_{N-1}) = x^{0j}(t_{N-1}), \qquad \text{si } i \neq j, \qquad i, j = 1, \ldots, m. \end{aligned} \tag{4.23}$$

La commande $u_i(t_{N-1})$ correspondant au vecteur $x(t_{N-1})$ donné sous la forme (4.23) aura les composantes suivantes:

$$u_i^s(t_{N-1}) = u^{0s} + \sum_j a_{sj}[x^j(t_{N-1}) - x^{0j}(t_{N-1})]$$

ou en vertu de (4.23)

$$u_i^s(t_{N-1}) = u^{0s} + a_{si}\delta^i.$$

La résolution de $m$ problèmes d'optimisation avec les conditions (4.23) nous donne $m$ vecteurs $u_k(t_{N-1})$ $(k = 1, \ldots, m)$. En posant $x(t_{N-1}) = x^1(t_{N-1})$ dans les équations (4.22), on obtient le système d'équations suivant:

$$\begin{aligned} u_1^1 &= u^{01} + a_{11}\delta^1, \\ u_1^2 &= u^{02} + a_{21}\delta^1, \\ &\ldots\ldots\ldots \\ u_1^h &= u^{0h} + a_{h1}\delta^1. \end{aligned} \tag{4.24}$$

Vu qu'on s'est donné $\delta^1$ et qu'on a calculé les $u_i^j$, le système d'équations (4.24) nous permet de déterminer les coefficients d'amplification $a_{j1}$:

$$a_{j1} = \frac{u_1^j - u^{0j}}{\delta^1}. \tag{4.25}$$

De façon analogue, en posant successivement $x(t_{N-1}) = x^2(t_{N-1})$, $x(t_{N-1}) = x^3(t_{N-1})$, etc., dans les équations (4.22), on définit les autres éléments

$$a_{ji} = \frac{u_i^j - u^{0j}}{\delta^i} \tag{4.26}$$

de la matrice $A_{N-1}$.

Passons maintenant à l'analyse de l'avant-dernier intervalle $t_{N-2}$. Dans cette étape on doit maximiser la fonctionnelle

$$I_{N-1} = \overline{F_{N-1}(x(t_{N-1}), u(t_{N-1}))} + \overline{F_N(x(t_N))}$$

sous les conditions

$$\begin{aligned} x(t_N) &= f_{N-1}(x(t_{N-1}), u(t_{N-1}), \xi(t_{N-1})), \\ x(t_{N-1}) &= f_{N-2}(x(t_{N-2}), u(t_{N-2}), \xi(t_{N-2})), \end{aligned} \tag{4.27}$$

où $x(t_{N-2}) = x_{N-2}$ est une quantité donnée. Dans ce problème on connaît la commande $u(t_{N-1})$, c'est la fonction (4.21) dont une approximation linéaire est donnée par les fonctions (4.25), (4.26). Donc, le problème d'optimisation de la fonctionnelle $I_{N-1}$ sous les conditions (4.27) ne contient qu'une inconnue: le vecteur $u(t_{N-2})$. Ce problème est par conséquent identique au précédent. Il risque d'être compliqué un peu plus par la présence de deux relations aux différences finies (4.15) au lieu d'une. Mais la dimension des deux problèmes, c'est-à-dire le nombre de quantités scalaires inconnues, est la même. La résolution du problème d'optimisation sous les conditions (4.27) est particulièrement simple si les fonctions $f_{N-1}$ et $f_{N-2}$ sont des fonctions linéaires de la commande, et $F_N$ et $F_{N-1}$, des fonctions linéaires de leurs arguments. Dans ce cas, le problème se ramène à un problème de programmation linéaire. Le calcul ne sera pas très long en raison de la faible dimension du problème.

Si les relations sont essentiellement non linéaires, on peut faire appel à la méthode de projection des gradients. Après les deux ou trois premiers pas, on aura intérêt à linéariser le problème au voisinage de l'approximation trouvée et ensuite à le résoudre comme un problème de programmation linéaire.

En étudiant la procédure de choix de la commande sur le dernier intervalle temporel, on a attiré l'attention sur le fait que le dispatcher tend à maximiser non pas la fonctionnelle de l'espérance mathématique, mais la fonctionnelle dans chaque cas concret. Ceci a permis d'utiliser l'information fournie par les observations de $\xi(t_{N-1})$. Dans ce cas aussi, le dispatcher connaît la réalisation de $\xi(t_{N-2})$ et pas seulement elle: il connaît toutes les valeurs de $\xi(t_N)$ pour $n = 0, 1, \ldots, N-2$. Ceci peut l'aider à procéder à une prédiction réaliste de la quantité $\xi(t_{N-1})$:

$$\xi_{N-1} = \Gamma_{N-1}(\xi(t_0), \quad \xi(t_1), \ldots, \xi(t_{N-2})).$$

Il est évident que la valeur prédite de $\xi_{N-1}$ ne coïncidera pas avec la réalisation de $\xi(t_{N-1})$ observée sur l'intervalle temporel suivant. Mais l'utilisation de la prédiction peut être mieux justifiée et fournir des résultats plus exacts que si $\xi(t_n)$ est supposée aléatoire.

La résolution de $m$ problèmes d'optimisation sous les conditions (4.23) pour le système (4.27) nous donne la synthèse de la commande sur l'intervalle de temps $t_{N-2}$:

$$u(t_{N-2}) = u_{N-2}(x(t_{N-2})). \tag{4.28}$$

La commande (4.28) a été déterminée sous la forme d'une fonction linéaire des écarts par rapport à la trajectoire optimale. On passe ensuite au choix de la commande sur l'intervalle de temps $t_{N-3}$; ceci nous conduit à un problème analogue au précédent qui consiste à maximiser la fonctionnelle $I_{N-2}$ mais déjà avec trois contraintes de la forme (4.27). La dimension du problème, c'est-à-dire le nombre des quantités scalaires inconnues, est toujours la même. Ce processus se poursuivra jusqu'à épuisement de tous les intervalles temporels.

Nous avons ainsi exposé une approche de la conception d'un système de gestion d'objets dynamiques du type cascade de barrages. Le schéma considéré mêlait des méthodes purement euristiques telles l'établissement des prédictions, et des méthodes numériques de programmation dynamique. Cette symbiose des méthodes euristiques et numériques fait recette dans de nombreux problèmes d'analyse des systèmes.

## § 5. Problèmes stochastiques

Les problèmes stochastiques jouent un rôle particulier en analyse des systèmes. Je ne pense pas exagérer en affirmant que tout problème d'optimisation bien « posé » est stochastique. Les facteurs aléatoires —

aléatoires par essence ou par ignorance des causes génératrices — se retrouvent dans pratiquement tout problème. Les problèmes déterministes ne sont rien d'autre que des problèmes stochastiques dans lesquels on néglige souvent le caractère aléatoire de tel ou tel paramètre.

Les problèmes stochastiques envisagés seront des problèmes d'utilisation de certaines ressources, c'est-à-dire de détermination d'une stratégie d'utilisation de ces ressources. Les situations les plus diverses peuvent se présenter. Dans certains cas il faudra trouver une stratégie déterministe bien définie, par exemple des paramètres d'une construction attaquée par des forces aléatoires, dans d'autres une stratégie qui est elle-même aléatoire (comme dans le problème de détection d'un objet se déplaçant aléatoirement dans un espace limité ou dans les problèmes de correction). A noter que les méthodes stochastiques peuvent être très utiles même dans la résolution de problèmes déterministes de dimension élevée. Les problèmes d'optimisation stochastique forment actuellement une branche développée des mathématiques appliquées. Ils ont fait l'objet de nombreux travaux remarquables et d'innombrables méthodes d'optimisation stochastique ont été élaborées pour les résoudre (cf. [2]).

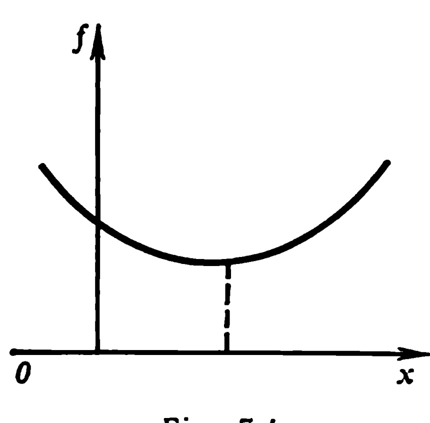


Fig. 7.4

L'exposé des problèmes de programmation stochastique outrepasse le cadre de cet ouvrage. Non seulement il exige beaucoup de place, mais il réclame une formation spéciale (la connaissance de nombreux chapitres spéciaux de la théorie des probabilités). Dans cet ouvrage on s'appesantira sur quelques exemples seulement illustrant la place et les possibilités des méthodes d'optimisation stochastique.

a) *Méthodes stochastiques de choix aléatoire.* Considérons un problème classique : la détermination du minimum d'une fonction convexe $f(x)$. La fonction de la figure 7.4 présente un seul minimum. Donc, si les courbes de niveau $f(x) = \text{const}$ ne présentent pas de vallées et sont assez lisses, la recherche du minimum ne soulève aucune difficulté fondamentale. Ce minimum peut être déterminé, par exemple, par la *méthode du gradient* dont le principe général est exprimé par la formule suivante

$$x_{k+1} = x_k - \alpha_k f'(x_k), \tag{5.1}$$

où $x_k$ est un point en lequel est calculée la fonction $f(x)$ ; $x_{k+1}$, un nouveau point ; $\alpha_k$, un scalaire appelé pas. Ce dernier est choisi de

telle sorte que

$$f(x_{k+1}) < f(x_k).$$

Si $\alpha_k$ est déterminé à partir de la condition

$$f(x_k - \alpha f'(x_k)) \Rightarrow \min_{\alpha},$$

alors la formule (5.1) décrit la *méthode de plus grande pente*. A chaque pas du processus (5.1) il faut calculer $n$ dérivées $\partial f/\partial x^i$. Supposons maintenant que la dimension $n$ du vecteur $x$ est assez grande et que le calcul de la fonction $f(x)$ est relativement complexe; la détermination numérique des dérivées

$$\frac{\partial f}{\partial x^i}(x_k) = \frac{f(x_k + \varepsilon e_i) - f(x_k)}{\varepsilon}, \quad i = 1, \ldots, n,$$

où $\varepsilon$ est un petit nombre et $e_i$ le vecteur unitaire du $i$-ème axe de coordonnées, peut nécessiter beaucoup de temps machine. Même si la méthode converge pour $\varepsilon \to 0$ et $k \to \infty$, sa réalisation est pratiquement impossible.

A la fin des années cinquante, au Centre de calcul de l'Académie des sciences d'U.R.S.S., nous avons appliqué la méthode de Monte-Carlo à la résolution de tels problèmes. Mais le choix aléatoire du vecteur $x$ eu égard à sa dimension élevée a conduit à des procédures très peu économiques. C'est pourquoi nous avons combiné la méthode du gradient et la méthode de Monte-Carlo. Nous nous sommes finalement arrêtés sur l'analogue suivant de la formule (5.1):

$$x_{k+1} = x_k - \alpha_k \frac{f(x_k + \beta\xi) - f(x_k)}{\beta} \xi, \tag{5.2}$$

où $\xi$ est un vecteur aléatoire, $\beta$ un petit scalaire dont le signe doit être pris tel que $f(x_k + \beta\xi) - f(x_k) > 0$. Le schéma (5.2) est une généralisation de la méthode séquentielle qui consiste à fixer les vecteurs unitaires $e_1$, $e_2$, etc., des axes de coordonnées et à faire un pas dans le sens de l'un de ces axes.

Si le problème n'est pas de dimension élevée, la méthode du gradient est toujours préférable à la méthode séquentielle. Mais l'efficacité relative de cette dernière croît avec la dimension du problème. L'efficacité d'une méthode numérique utilisant un nombre élevé d'itérations se juge sur deux caractéristiques: le nombre d'itérations et la durée d'une itération. Plus la dimension croît et plus la durée d'une itération fait pencher la balance en faveur de la méthode séquentielle. Certes, lorsque la dimension du problème croissait, le nombre d'itérations augmentait moins vite pour la méthode du gradient que pour la méthode séquentielle, mais dans l'ensemble nous avons constaté une amélioration de l'efficacité de cette dernière.

Le mode de choix du pas $\alpha_k$ a joué un certain rôle. Pour la méthode séquentielle on peut proposer aussi un moyen de choisir le pas qui est identique à celui de la méthode de plus grande pente. Mais le choix du pas à partir de la condition

$$f(x_{k+1}) \Rightarrow \min_{\alpha} \tag{5.3}$$

accroît considérablement la durée d'une itération. C'est pourquoi nous avons remplacé la condition (5.3) par la condition plus simple

$$f(x_{k+1}) < f(x_k). \tag{5.4}$$

La condition (5.4) peut toujours être réalisée si l'on choisit $\alpha$ suffisamment petit. Les expériences effectuées avec la méthode séquentielle étaient une étape naturelle de passage aux méthodes par tâtonnements: il restait à remplacer le vecteur $e_i$ par un vecteur aléatoire $\xi$. Il se trouve que dans les problèmes de grande dimension ($n$ est de l'ordre de quelques dizaines) le passage à un choix aléatoire du sens de $\xi$ (au moyen d'un générateur de nombres aléatoires) rend le schéma (5.2) plus performant. Ce fait a été attesté par d'innombrables expériences. Plus tard, lorsque la théorie de la programmation stochastique fut élaborée, on s'aperçut que le schéma (5.2) était un cas très particulier de la méthode des quasi-gradients stochastiques proposée par Yu. Ermoliev. De la théorie générale qu'il a développée, il ressort que la méthode basée sur la procédure itérative (5.2) converge pour les fonctions convexes si seulement sont remplies les deux conditions suivantes:

$$\alpha_k \to 0, \quad \sum \alpha_k = \infty.$$

(Pour plus de détails voir l'ouvrage [2] déjà cité.) Donc, la méthode de choix aléatoire qui somme toute est une méthode stochastique, se prête bien à la résolution de problèmes déterministes d'optimisation.

Les expériences effectuées au Centre de calcul de l'Académie des sciences d'U.R.S.S. datent de l'époque où la méthode de choix aléatoire commençait à peine d'être utilisée en tant que cas particulier des méthodes stochastiques d'optimisation. Depuis, ces méthodes ont fait fortune et occupent une place importante parmi les méthodes numériques de résolution des problèmes d'optimisation (cf. par exemple [11]).

Si dans les problèmes déterministes de recherche d'un extrémum local (ou dans les problèmes à un seul extrémum) l'emploi des méthodes stochastiques n'est justifié que si les problèmes sont de dimension très élevée ou si le calcul des valeurs des fonctions demande beaucoup de temps machine, dans les problèmes de détermination de l'extrémum global, les méthodes stochastiques sont hors pair.

En effet, toute méthode déterministe est basée sur l'analyse des propriétés locales de la fonction étudiée. On calcule la valeur de la fonction $f(x)$ en un point quelconque $x$, puis on étudie le voisinage de $x$ par une méthode possible. Si l'on trouve des points $x^1$ vérifiant les contraintes et tels que $f(x^1) < f(x)$, on passe à l'itération suivante. Mais ce procédé n'est bon que pour trouver l'extrémum local. Si l'on veut déterminer l'extrémum global, il faut des méthodes basées sur un autre principe: en l'occurrence des méthodes stochastiques. Ces dernières années, on accorde une attention accrue à l'application de ces méthodes à la recherche de l'extrémum global (cf. par exemple [12]).

b) *Problèmes stochastiques.* Jusqu'ici nous avons parlé de l'application des méthodes stochastiques à la résolution de problèmes déterministes. Arrêtons-nous maintenant sur les problèmes essentiellement stochastiques qui font intervenir des fonctions ou des quantités aléatoires. L'archétype de ces problèmes est le problème de recherche du minimum de l'espérance mathématique d'une fonction

$$\overline{f(x, \xi)} \Rightarrow \min_x \tag{5.5}$$

sous les contraintes

$$\overline{G(x, \xi)} \leqslant 0, \tag{5.6}$$

$$\Gamma(x, \xi) \leqslant 0 \qquad \forall \xi. \tag{5.7}$$

L'espérance mathématique (5.5) est prise sur la variable $\xi$ qui est supposée être une fonction ou une quantité aléatoire et la minimisation est effectuée sur la variable $x$ qui est un élément d'un espace fonctionnel.

Les contraintes du problème sont divisées en deux groupes. Le premier groupe de contraintes (5.6) a la forme d'une espérance mathématique par rapport à $\xi$. $G$ est un opérateur qui peut avoir une structure assez complexe. Ce peut être, en particulier, un système d'équations différentielles ou aux différences qui associe aux fonctions $x(t)$ et $\xi(t)$ un système de fonctionnelles terminales $\overline{F_i(x(T))}$.

Les contraintes (5.7) sont déterministes. Par exemple, quelles que soient les perturbations aléatoires attaquant un avion (coups de vent ou fluctuations de la poussée du moteur), les angles de rotation des gouvernes ne peuvent varier que dans une plage définie qui est une caractéristique de cet avion. De façon exactement analogue, dans le problème de mise sur orbite d'un engin spatial, la fonctionnelle sera la précision du tir, c'est-à-dire une fonctionnelle de type variance

$$J = \overline{((x(T) - x_T), R(x(T) - x_T))}, \tag{5.8}$$

et les contraintes seront déterministes: la poussée propulsive $p(t)$ et l'angle de rotation $\alpha(t)$ des gouvernes doivent satisfaire des inégalités de la forme

$$0 \leqslant p(t) \leqslant p^+, \qquad \alpha^- \leqslant \alpha(t) \leqslant \alpha^+ \qquad \forall t.$$

Le caractère des contraintes (5.6), (5.7) et la structure de la fonctionnelle (5.5) compliquent énormément la recherche de l'élément $x(t)$ minimisant la fonctionnelle, la principale difficulté étant soulevée par le calcul des expressions (5.5), (5.6) même si l'élément $x(t)$ est donné. En effet, si nous avions pu calculer facilement les espérances mathématiques (5.5) et (5.6), le problème se serait ramené à un simple problème déterministe d'optimisation, car dans ce cas on aurait pu le mettre sous la forme

$$\overline{f(x, \xi)} = f^*(x) \Rightarrow \min, \tag{5.5'}$$

$$\overline{G(x, \xi)} = G^*(x) \leqslant 0. \tag{5.6'}$$

Donc, toutes les particularités et les difficultés du problème (5.5), (5.6), (5.7) résident en premier lieu dans le calcul des espérances mathématiques. La condition « déterministe » (5.7) ne fait qu'aggraver ces difficultés.

Lorsque dans les années soixante, les analystes furent confrontés à des problèmes de type (5.5), (5.6), (5.7), leur premier réflexe fut de les ramener à des problèmes déterministes (par le calcul des espérances mathématiques) pour leur appliquer ensuite les méthodes d'optimisation bien élaborées. Ces tentatives n'ont pas abouti à la création d'un appareil mathématique assez universel bien qu'elles aient permis de résoudre avec succès de nombreux problèmes d'application.

Les méthodes euristiques de résolution des problèmes stochastiques se sont largement développées. Arrêtons-nous sur l'une d'elles.

Supposons que le problème (5.5), (5.6), (5.7) puisse être résolu assez facilement pour des valeurs fixes du vecteur aléatoire $\xi$. La procédure suivante semble la plus logique. Le générateur de nombres aléatoires nous délivre une valeur $\xi = \xi_1$. Résolvons par une méthode déterministe le problème déterministe

$$\begin{aligned} &f(x, \xi_1) \Rightarrow \min, \\ &G(x, \xi_1) \leqslant 0, \\ &\Gamma(x, \xi_1) \leqslant 0 \end{aligned} \tag{5.9}$$

et désignons sa solution par $x_1$. Faisons délivrer par le générateur une autre valeur, $\xi_2$, et résolvons un nouveau problème d'optimisation (5.9). Désignons sa solution par $x_2$. En répétant cette procédure $N$ fois, on obtient les valeurs

$$x_1, x_2, \ldots, x_N$$

et pour solution on peut prendre

$$x^* = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{5.10}$$

Donc, au lieu de résoudre le problème $\min_x \overline{f(x, \xi)}$, nous avons trouvé la solution $x^*$ du problème $x^* = \overline{\arg \min_x f(x, \xi)}$. Mais dans le cas général l'opération min et le calcul de l'espéranee mathématique ne commutent pas et l'expression (5.10) n'est pas solution du problème d'optimisation primitif (5.5), (5.6), (5.7).

Elucidons cette situation sur un exemple simple. Soit le problème

$$\overline{f(x, \xi)} = \overline{x^2 + \xi x} \Rightarrow \min_x \tag{5.11}$$

sous réserve que $\xi$ soit une quantité aléatoire centrée, c'est-à-dire que $\bar{\xi} = 0$. Le vecteur $x$ étant déterministe, on a

$$\overline{f(x, \xi)} = f^*(x) = x^2 \text{ et } \min_x f^*(x) = 0.$$

Effectuons les calculs d'après le procédé décrit. Le minimum de la fonction $f(x, \xi) = x^2 + \xi x$ pour $\xi$ fixe étant réalisé pour $x = -\xi/2$, on a

$$\min_x f(x, \xi) = -\xi^2/4,$$

donc l'espérance mathématique de cette quantité est de la forme

$$\overline{\min_x f(x, \xi)} = -\frac{1}{4}\sigma^2(\xi),$$

c'est-à-dire est proportionnelle à la variance. Donc, plus la variance de la variable aléatoire $\xi$ sera grande et plus l'écart entre le résultat obtenu par la méthode euristique et le résultat exact sera grand. Il est tout aussi facile d'exhiber des exemples dans lesquels les résultats acquis par les deux méthodes sont confondus. (Signalons que dans l'exemple (5.11) la variable aléatoire figure linéairement dans la fonction.) Mais nous ne disposons d'aucun critère rigoureux nous permettant de juger de l'adéquation (ou de la précision) de la méthode euristique décrite. Ceci n'empêche que le procédé qui consiste à ramener la minimisation de l'espérance mathématique d'une fonction au calcul de l'espérance mathématique de ses extrémums est assez répandu dans les problèmes d'ingénieur et les problèmes économiques. Dans le cas général cette méthode ne nous fournit certes pas de commande optimale ou de système optimal de paramètres.

Mais la commande donnée par cette méthode euristique est souvent admissible, c'est-à-dire vérifiant les contraintes (tout au moins les plus compliquées d'entre elles: les contraines déterministes (5.7)).

J'ai déjà signalé à plusieurs reprises que le choix de la méthode relevait toujours d'un compromis: il faut concilier deux impératifs: la précision du résultat et son prix de revient. Donc le fait que la réduction du problème de minimisation de l'espérance mathématique au calcul de l'espérance mathématique du minimum permet de réduire de plusieurs fois le temps machine est un argument de poids en faveur de la méthode euristique décrite. Enfin, il est question ici du choix de la forme du critère, or on sait qu'en la matière l'arbitraire est grand. A noter encore que la solution acquise par la méthode euristique décrite peut servir de bonne première approximation pour construire une théorie des perturbations ou d'autres procédures itératives.

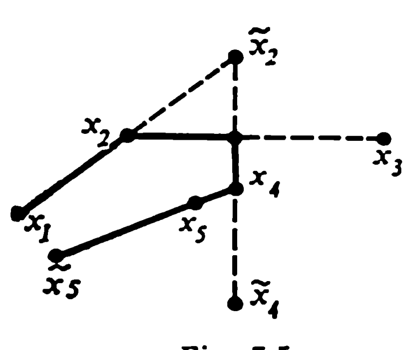


Fig. 7.5

c) *Généralisation stochastique de la méthode du gradient.* Comme déjà signalé, il existe plusieurs méthodes numériques assez efficaces pour la résolution des diverses classes de problèmes d'optimisation: la méthode d'approximation stochastique, la méthode des quasi-gradients stochastiques, etc. Dans ce numéro, on se propose d'étudier une variante de ces méthodes qui est un cas particulier de la méthode générale proposée par Yu. Ermoliev.

Considérons de nouveau le problème (5.5), (5.6), (5.7) en ne nourrissant aucun espoir de calculer l'espérance mathématique de la fonction $f(x, \xi)$ pour $x$ fixe. Considérons par ailleurs une classe de problèmes pour lesquels le calcul de l'extrémum de la fonction (5.5) sous les contraintes (5.6) et (5.7) et sous réserve que le paramètre aléatoire $\xi$ soit fixe ne donne lieu à aucune complication.

Dans le numéro précédent nous avons essayé de tirer un parti immédiat de cette circonstance. Ici nous adoptons une autre tactique: nous allons utiliser cette particularité du problème étudié pour construire une procédure de descente. Faisons délivrer une valeur $\xi_1$ au générateur de nombres aléatoires. Résolvons le problème d'optimisation (5.9) pour la valeur fixe $\xi_1$ du vecteur aléatoire. Nous obtenons ainsi le vecteur $x = x_1$ (fig. 7.5). Faisons de nouveau délivrer une valeur aléatoire $\xi_2$ par le générateur de nombres, resolvons le problème d'optimisation (5.9) pour $\xi = \xi_2$ et trouvons une nouvelle valeur $x = x_2$.

Pour itération suivante nous prendrons le vecteur

$$\tilde{x}_2 = x_1 + \alpha_1 (x_2 - x_1).$$

Déterminons ensuite la valeur $x = x_3$ à l'aide de la procédure initiale et formons l'itération suivante

$$\tilde{x}_3 = \tilde{x}_2 + \alpha_2 (x_3 - \tilde{x}_2),$$

et ainsi de suite.

Cette procédure qui de prime abord semble être euristique converge effectivement vers la solution exacte du problème primitif (5.5), (5.6), (5.7). Pour être convergente, la procédure itérative

$$\tilde{x}_{k+1} = \tilde{x}_k + \alpha_k (x_{k+1} - \tilde{x}_k), \tag{5.12}$$

où $x_{k+1}$ est la solution du problème d'optimisation (5.9) dans lequel on attribue au vecteur $\xi$ la valeur $\xi_{k+1}$ délivrée par le générateur de nombres aléatoires, doit satisfaire une série de conditions, dont les plus importantes *) portent sur le choix du pas $\alpha_k$:

$$\alpha_k \underset{k\to\infty}{\to} 0, \quad \sum \alpha_k = \infty, \quad \sum \alpha_k^2 < \infty. \tag{5.13}$$

La procédure itérative proposée n'est pas relaxative. Cela signifie que dans le cas général les inégalités $f^* (\tilde{x}_{k+1}) < f^* (\tilde{x}_k)$, où $f^* (\tilde{x}) = \overline{f (x, \xi)}$, ne sont pas réalisées, c'est-à-dire que la procédure (5.12) nous rapprochera et nous éloignera de la valeur cherchée du vecteur $x$. Ceci distingue essentiellement les méthodes stochastiques des méthodes déterministes les plus couramment utilisées.

---

*) Les conditions de convergence de la méthode d'approximation stochastique (5.13) ont probablement été établies pour la première fois par D v o r e t s k i (le théorème de Dvoretski est accessible dans [15]).

CHAPITRE VIII

# QUELQUES PROBLÈMES DE L'AUTOMATISATION DE L'ÉTUDE DES PROJETS

## § 1. Considérations générales

Dans ce chapitre le terme « automatisation de l'étude des projets » est pris dans une acception très large: il désigne une nouvelle technique d'étude s'appuyant sur les idées et les méthodes de l'analyse des systèmes. Les objets projetés peuvent être des systèmes techniques (avion, vaisseau, engin spatial) ou des systèmes économiques (exploitation de gisements de pétrole ou de gaz, utilisation des bassins fluviaux, etc.).

Les techniques utilisées et surtout celles en cours de création, les technologies, les rapports de production, les voies de communication se compliquent sans cesse (d'une manière exponentielle). Cette tendance est caractéristique de l'époque du développement industriel qualifiée communément d'époque de la révolution scientifique et technique *).

Les machines construites par les ingénieurs concrétisent de plus en plus les acquis de sciences connexes. Signe des temps, de nombreux secteurs de pointe tels l'astronautique, l'aéronautique, l'énergétique, etc., réclament la participation de la radio-électronique, de la théorie des processus thermiques, de la dynamique des gaz, etc. La brusque complication des rapports de production et des technologies, l'implantation de nouveaux matériaux rendent très difficile la tâche du projeteur à telle enseigne que sa façon de procéder est fondamentalement différente de celle de son grand-père.

De l'avis des spécialistes, la complexité des constructions mécaniques a été multipliée par six au cours des trente dernières années. Les méthodes utilisées par les constructeurs et les procédures d'étude des projets ont-elles suivi ce rythme? Très peu. C'est la raison pour laquelle la complication des constructions projetées bouscule les principes traditionnels d'étude des projets qui supposaient toujours au constructeur en chef la maîtrise totale de la construction projetée.

Il est évident que cette thèse sera dépassée tôt ou tard, étant donné

*) A mon avis le terme révolution scientifique et technique n'est pas très heureux. Après le brusque changement des technologies qui a eu lieu juste après la seconde guerre mondiale, les rythmes du progrès scientifique et technique n'ont pas baissé jusqu'à leur niveau d'avant-guerre, mais bien au contraire ont continué à croître. La révolution scientifique et technique est une étape naturelle du développement de la technique et de l'économie.

que les facultés physiologiques de l'homme sont limitées et que la complexité des constructions ne cesse de croître. Ces dernières décennies le constructeur en chef ou le chef d'un projet peuvent de moins en moins intervenir efficacement dans l'étude du projet. De créateur il se transforme, en mettant les choses au mieux, en bon administrateur.

Ceci pose avec acuité le problème de la promotion de nouvelles techniques d'étude: le problème de l'automatisation de l'étude des projets. Ces techniques devront utiliser largement les méthodes modernes de traitement et de représentation de l'information sous une forme qui permette au constructeur ou au projeteur de donner la pleine mesure de son talent. Ce problème est au cœur des préoccupations de tous les pays industrialisés qui construisent des engins techniques complexes ou étudient des projets économiques grandioses. L'automatisation de l'étude des projets devient progressivement le domaine de prédilection des ordinateurs et des méthodes interdisciplinaires d'étude de processus de nature physique différente.

En traitant les problèmes d'organisation d'une expertise, on a constaté que pour apporter une réponse à un problème complexe il fallait nécessairement le décomposer en sous-problèmes plus simples. De même un projet complexe doit être décomposé en sous-projets qui seront confiés à divers constructeurs et projeteurs. Cette situation ne date pas d'hier: l'étude du projet d'un avion ou d'un complexe industriel est toujours conduite par une équipe dont chaque membre remplit des fonctions bien définies.

Mais la décomposition d'un problème implique nécessairement le processus inverse: le remembrement et la coordination des caractéristiques des divers éléments du système et une synthèse qui permet au constructeur de se faire une idée globale de la construction, d'apprécier ses diverses qualités et son degré de fidélité.

La décomposition de l'étude des projets n'a initialement soulevé aucun problème. Prenons par exemple le cas d'un avion. Les études de cellule et de moteur sont fondamentalement différentes. Les calculs aérodynamiques et de résistance sont confiés à des experts de spécialité différente, etc. Cette situation prévalait partout. Et les formes traditionnelles de division du travail s'instaurèrent progressivement partout.

Pendant longtemps la synthèse du projet n'apporta pas de complications spéciales: les méthodes d'étude se perfectionnaient à mesure que se compliquaient les constructions projetées. Mais ces méthodes traditionnelles d'étude commencèrent à battre de l'aile.

Tout d'abord les délais d'étude s'allongèrent de façon inadmissible. Mais ce n'était qu'un demi-mal. Le pire c'est qu'au banc d'essai on voyait arriver des dispositifs qui correspondaient de moins en moins aux desseins du constructeur qui n'avait pas les moyens de vérifier assez bien le degré de fidélité de son prototype. Ce qui entraî-

28—01523

ne d'inévitables réfections, un renchérissement du dispositif construit et un retard de plusieurs années. Dans les faits cela se traduit par la mise en service de dispositifs (ou de technologies) obsolètes qui accusent un retard au niveau technique de dix à vingt ans.

L'analyse de ces phénomènes nous convainc que les principales difficultés proviennent de la synthèse et de la coordination de toutes les particularités de la future construction. Ces difficultés croissent en raison exponentielle de la dimension, c'est-à-dire du nombre de paramètres qui définissent la construction. L'habileté des projeteurs n'est pas en cause ici : la technologie traditionnelle est en principe incapable de venir à bout des difficultés croissantes du projet et il faut la changer.

L'émergence et la maturation des conceptions de l'automatisation de l'étude des projets se sont déroulées en gros de la manière suivante.

On a d'abord commencé par automatiser le dessin industriel qui constitue la partie la plus laborieuse de tout projet. Des dessinateurs automatiques ont été construits. Si les dépenses ont été très vite amorties, ces dessinateurs n'ont et ne pouvaient apporter aucune innovation susceptible d'améliorer ou d'accélérer le projet.

On a parallèlement largement implanté les méthodes de calcul sur ordinateurs dans les calculs d'ingénieurs (par exemple de résistance, hydrauliques, aérodynamiques, etc.). Ces méthodes, même si elles ont sensiblement amélioré les diverses procédures de calcul des projets, réduit au minimum les erreurs possibles et contribué à élever la « culture générale » du projeteur, n'ont pas entraîné une réduction sensible des délais de réalisation des projets.

A dire vrai, de grands espoirs ont été placés dans l'utilisation des ordinateurs pour les calculs d'ingénieurs et de planning. Mais ces espoirs n'ont pas été justifiés en grande partie. La faute n'incombe pas aux ordinateurs, mais aux techniciens qui les maîtrisent encore assez mal. A telle enseigne que pendant longtemps les ordinateurs ont joué le rôle de grands arithmomètres *). Ils ont permis de résoudre certains problèmes d'ingénieurs plus rapidement et plus exactement, mais ils étaient encore incapables d'influer sérieusement sur le déroulement du projet, de faire avancer la fin des travaux.

L'étape suivante a consisté à créer des postes de travail automatiques pour les constructeurs. Un important pas en avant avait été accompli. Ces postes communiquaient directement avec l'ordinateur qui remplaçait ainsi le té traditionnel, l'arithmomètre ou la règle à calcul ; des écrans vidéo élémentaires permettaient au constructeur de dialoguer avec l'ordinateur. L'idée des postes de tra-

*) Le rôle et les fonctions des ordinateurs sont ainsi appréhendés par une grande partie d'ingénieurs, d'économistes et même de mathématiciens.

vail automatiques a émergé à la fin des années soixante en même temps que les systèmes à temps partagé. Bien des espoirs ont été associés à leur implantation. Et s'ils n'ont pas tous été fondés, il n'en reste pas moins que toutes les dépenses occasionnées ont été entièrement justifiées. Le seul fait qu'ils ont contribué à élever le niveau de qualification du constructeur et à élargir l'horizon de ses connaissances valait la peine. Une autre conséquence importante de la création des postes automatiques fut l'instauration d'un dialogue entre l'ordinateur et le constructeur. L'automatisation de l'étude des projets entrait ainsi dans une nouvelle phase de son développement. Le domaine des ordinateurs était jusqu'ici le fief du mathématicien, c'est ce dernier qui résolvait les problèmes pour le compte du constructeur. Désormais le constructeur était à même de se mettre à un pupitre. Il est évident que la qualité des projets fut améliorée.

Mais l'automatisation des postes de travail qui s'est opérée dans de nombreux pays au début des années soixante-dix n'a pas apporté une solution au problème central. Entre l'émergence d'une idée et sa réalisation il s'écoulait beaucoup de temps. Les dispositifs (ou leurs modèles) envoyés au banc d'essais nécessitaient d'innombrables et difficiles remaniements au cours des essais et parfois même un refaçonnage.

Dans les cas où il est impossible de procéder à des essais, par exemple lors de l'implantation d'un complexe industriel, les défauts du projet (défauts qu'il est impossible de mettre en évidence sur le papier) peuvent se solder par des catastrophes. Et il ne pouvait en être autrement, car le poste de constructeur n'est, somme toute, qu'une partie du système d'étude du projet.

On sentait la nécessité d'un système coordonné d'étude des projets comprenant un sous-système de programmes pour les calculs d'ingénieurs, des postes de travail automatiques, des procédures de dialogue variées et, bien sûr, des travaux graphiques automatiques. De nombreux pays ont mis actuellement sur pied d'intenses travaux pour la construction et l'implantation de tels systèmes.

Il est encore prématuré de dresser un bilan, de parler de l'efficacité de ces systèmes. Les espérances fondées sur l'implantation de ces systèmes expliquent que l'automatisation de l'étude des projets soient en vedette. Qu'il me soit permis à ce propos de mettre en garde les esprits enthousiastes: la mise en place de systèmes automatiques d'étude est un travail de longue haleine.

Tout d'abord ils impliquent des ordinateurs très perfectionnés et un système d'utilisation collective développé. Les systèmes d'étude tels qu'ils sont conçus aujourd'hui nécessitent un usage collectif des banques de données, des modèles et des programmes. Leur exploitation réclame un grand nombre de disques magnétiques (d'une très grande capacité de mémorisation), des terminaux spéciaux, etc. Enfin la création d'un logiciel spécial demande de nombreuses an-

nées d'efforts acharnés à plusieurs équipes hautement qualifiées: tout système d'automatisation de l'étude des projets est aujourd'hui un système de simulation très sophistiqué. C'est pourquoi l'apparition de systèmes valables est pour le milieu de la prochaine décennie.

Cette prévision appelle une remarque importante: les systèmes doivent être conçus et implantés graduellement à mesure que les divers éléments seront opérationnels, c'est-à-dire que l'exploitation des sous-systèmes devra commencer bien avant que le système complet soit achevé. C'est un principe important qui permettra d'économiser plusieurs millions de roubles. Mais sa réalisation réclame non seulement un logiciel spécial, mais une organisation spéciale du travail des utilisateurs: les constructeurs et les projeteurs.

Voyons maintenant le problème du dialogue. Tout le monde s'accorde aujourd'hui à penser que le système d'automatisation d'étude des projets est un système spécial de dialogue et que le dialogue entre l'homme et l'ordinateur est l'aspect le plus important de l'étude des projets. Mais malheureusement d'aucuns estiment que l'organisation du dialogue n'est pas un problème scientifique et qu'elle se ramène avant tout à la résolution de questions purement techniques de création de terminaux spéciaux et d'un bon logiciel (paquets de programmes pour la résolution de problèmes d'ingénieurs). C'est une erreur monumentale. Si elle n'est pas écartée à temps, les systèmes créés risquent d'apporter bien des désillusions.

Certes sans ordinateurs perfectionnés il ne peut être question de bons systèmes automatiques d'étude. Ceci concerne encore plus le logiciel. Mais en fin de compte ce ne sont que des instruments techniques qu'il faudra apprendre à manier. Plus l'instrument est sophistiqué, plus il est difficile de le manipuler et plus la formation de l'utilisateur doit être élevée. L'expérience de construction et d'exploitation des ordinateurs le confirme totalement. Je suis convaincu pour ma part que le niveau des ingénieurs qui conçoivent les ordinateurs et leurs accessoires est bien plus élevé que celui des utilisateurs. Ceci montre qu'une infime partie seulement des possibilités des ordinateurs est mise à profit.

Analysons quelques difficultés rencontrées par les spécialistes chargés des problèmes de l'utilisation des ordinateurs dans l'étude des projets. Supposons que nous disposons d'un ordinateur idéal pour déterminer les caractéristiques de la construction projetée. Mais l'efficacité de tout dispositif complexe, qu'il soit industriel, technique ou économique, dépend de milliers de facteurs. Nous n'allons tout de même pas envisager toutes les combinaisons possibles de ces facteurs! Tout ordinateur, même d'une puissance illimitée, serait trop faible pour cela. La première des choses donc est d'élaborer une méthode économique de recherche de la meilleure variante.

En somme, il est nécessaire d'élaborer un système spécial de

règles et d'algorithmes qui servira de base à la nouvelle technique d'automatisation d'étude des projets. Sans cette nouvelle technique, les systèmes d'automatisation d'étude des projets du genre postes de travail automatiques seront des instruments utiles qui perfectionneront certes l'étude des projets, mais qui n'apporteront probablement pas des changements de nature à l'améliorer.

Ainsi, le problème central de la conception des systèmes d'automatisation de l'étude est l'élaboration d'une nouvelle technique d'étude. L'élaboration de cette technique est sur le métier dans de nombreux bureaux qui mettent au point et utilisent de tels systèmes. Le recyclage des constructeurs et des projeteurs est une partie intégrante de cette nouvelle technique. Son élaboration est encore pour une grande part affaire d'intuition et n'a pas assez mobilisé les efforts conjugués des ingénieurs, des projeteurs et des mathématiciens.

Quelles seront les grandes lignes de cette action ? L'auteur de cet ouvrage a un jour posé cette question à l'un des plus célèbres constructeurs d'avions soviétiques, feu P. Soukhoï. La réponse fut sans équivoque : « Vous les mathématiciens de la machine, vous devez m'aider, moi, le constructeur en chef. Vous devez élaborer avec moi une méthode d'étude des projets qui permette au stade initial déjà de choisir assez correctement les principaux paramètres de la construction et d'évaluer les diverses caractéristiques de son efficacité et de contrôler ensuite leurs variations de manière à envoyer au banc d'essais un modèle qui n'exige pas de finition. L'erreur commise par le constructeur à l'étape initiale du projet ne peut être réparée ni par des calculs d'ingénieur parfaits ni par des dessinateurs automatiques. »

Je pense que ces principes sont fondamentaux. Leur mise en œuvre peut effectivement contribuer de façon décisive à améliorer l'état des choses dans l'étude des projets. Mais qu'on ne s'y trompe pas, c'est un problème complexe qui implique la création d'un appareil spécial.

Faisons maintenant quelques remarques sur la « théorie » des procédures non formelles et sur son application à l'étude des projets de dispositifs techniques complexes. La construction d'un avion, d'un ordinateur, etc., est d'abord un acte de création qui en tant que tel ne peut jamais être entièrement formalisé. Nous érigerons ce fait en postulat. A noter que de nombreux spécialistes aussi bien soviétiques qu'étrangers estiment que l'acte de création dans l'étude des projets peut être en grande partie remplacé par un système spécial de traitement du matériel statistique. Le traitement statistique des paramètres des dispositifs existants (ou éventuels) est manifestement très important et ne doit en aucune façon être déprécié. Mais ce traitement ne suffit pas à lui seul. L'utilisation du seul matériel statistique permet de construire un dispositif semblable (ou pro-

che) de dispositifs déjà existants. En effet, les constructions originales concrétisent toujours des idées hardies qui sortent des sentiers battus. Il est donc impossible de les acquérir sur la base de données statistiques: ceci constituera le deuxième postulat. Mais une fois ce postulat (c'est-à-dire l'impossibilité d'une formalisation totale) adopté, il faut appréhender la place et la portée des méthodes formelles, voir comment elles pourraient être utiles au constructeur et comment elles pourraient être combinées aux procédures non formelles.

L'étude et l'organisation des procédures euristiques mobilisent l'attention d'un grand nombre de spécialistes qui leur ont consacré une énorme quantité de travaux. Une certaine compréhension des principes qui les régissent s'est dégagée. Ceci a été mentionné au chapitre précédent. Yu. Hermeyer [3, 4] a apporté une importante contribution à ce domaine en développant un nouveau système de conceptions sur le contenu des procédures non formelles et des résultats acquis grâce à elles. Yu. Jouravlev [38] a proposé un schéma général de formalisation de l'ensemble des procédures euristiques et de choix de la meilleure d'entre elles. G. Pospélov [62] et V. Glouchkov [30] ont exhibé des exemples d'application réussie des principes de décomposition dans la résolution de problèmes complexes d'évaluation des perspectives de développement. Bref, à ce jour on a accumulé suffisamment de connaissances pour construire des procédures euristiques pas seulement intuitivement.

Le principe de la décomposition qui a longuement été abordé au chapitre précédent est très important dans l'étude des projets de constructions complexes. A noter que ce principe est à la base de toutes les techniques d'étude des projets, pour peu que ces projets soient assez complexes. Et cela coule de source, car tout constructeur, si doué qu'il fût, ne peut manipuler qu'un volume assez limité d'informations (de paramètres, de critères, etc.). Donc, le mécanisme d'étude doit être conçu de manière à ne confier que des tâches relativement simples à chaque participant de la procédure. Ceci est le gage de la réussite et nous en voulons pour preuve les exemples du chapitre précédent. Signalons encore un fait: la décomposition doit être adaptée à la synthèse. Malheureusement il n'y a pas de recommandations générales en la matière. Dans chaque cas les projeteurs doivent utiliser des procédés originaux pour étudier et projeter la structure hiérarchique des problèmes et leurs interactions.

Illustrons ceci sur l'étude du projet d'un avion quoique une bonne part de ce qui va suivre est valable pour toute construction complexe. A la tête de la hiérarchie considérée se trouve le constructeur en chef dont la tâche est de choisir les paramètres qui fourniront la solution des problèmes posés par le client. Dans le cas d'un avion de transport de passagers, le client est le Ministère de l'aviation civile. Celui-ci veut un avion capable d'atterrir sur un aérodrome en terre et qui

soit meilleur que le YAK-40 et le AN-24 actuellement exploités. Dans le cas d'un avion de guerre, le client (en l'occurrence le Ministère de la défense) veut un chasseur qui surpasse les modèles précédents (en particulier, qui soit capable de les détruire dans un combat). Le problème doit être posé en ces termes: cette position est naturelle. Aucun mathématicien, aucun constructeur n'est en mesure de composer une fonctionnelle $F(x)$ dépendant de tous les paramètres $x$ de l'avion et dont la maximisation fournisse la solution du problème. En réalité, la fonctionnelle $F(x)$ ne dépend pas uniquement des paramètres $x$ de l'avion, elle dépend aussi d'un grand nombre de facteurs indéterminés $y \in Y$ caractérisant le milieu dans lequel évoluera l'avion (font partie des facteurs $y$ les paramètres de construction des contre-mesures de l'adversaire). Donc $F = F(x, y)$.

Pourtant cette fonctionnelle existe objectivement et assez souvent, en tout cas chaque fois que le niveau technique de l'industrie et le niveau des connaissances des lois régissant le fonctionnement du dispositif permettent en principe de le fabriquer, c'est-à-dire si le problème posé admet une solution. En effet, si deux prototypes sont proposés, alors l'expert (le client ou le constructeur en chef) peut choisir le meilleur après des tests complets. Les conditions où cette procédure est possible seront appelées critères de compétence. Ils disent que nous avons (plus exactement, le client ou le constructeur) une idée de ce qui est « meilleur ». Ce n'est qu'à cette condition qu'on peut poser le problème et choisir la meilleure construction. Si le client ou le constructeur ne peuvent pas choisir le meilleur de deux modèles, c'est que ou bien ces modèles sont équivalents, ou bien les critères de compétence ne sont pas remplis par le décideur, et dans un cas comme dans l'autre le mathématicien ne peut être d'aucun secours.

Donc, le constructeur est confronté au problème inextricable du choix des paramètres $x$ maximisant une fonctionnelle $F(x, y)$ qui non seulement ne peut être explicitée mais ne peut être formellement décrite.

Nous allons néanmoins tenter de nous représenter les voies possibles de résolution de ce problème. La simulation nous semble bien indiquée. Voyons les deux types d'avions envisagés plus haut; commençons par le chasseur. Supposons que nous avons mis au point un système simulant le combat de deux chasseurs. Introduisons dans l'ordinateur les paramètres du chasseur projeté et ceux du chasseur réel et simulons un combat entre eux. Les informations recueillies nous permettent de définir le « meilleur ». Comme il est question de la maîtrise de l'air, le meilleur est celui qui aura remporté le plus grand nombre de combats.

La situation est bien plus compliquée pour l'avion civil en ce sens qu'on ne dispose pas d'un critère évident de qualité. Mais par une série de simulations on offre à l'expert, sous réserve qu'il satis-

fasse au critère de compétence, la possibilité de choisir la meilleure variante.

Donc, le système de simulation permet en principe de comparer les variantes et de choisir la meilleure. Or cela signifie qu'il est possible de chercher le maximum d'une fonctionnelle dont on ignore la forme explicite. Mais ce n'est qu'une possibilité de « principe » de se servir du système de simulation comme d'un instrument d'optimisation. Voyons cette question de plus près. La première option de l'avion n'est jamais assez bonne quels que soient le talent et l'expérience du constructeur. Celui-ci doit résoudre un problème classique d'amélioration des paramètres, c'est-à-dire de remplacement d'un système de paramètres par un autre. Le système de simulation s'il n'est pas assez « intelligent », c'est-à-dire s'il n'est pas équipé de procédures spéciales ne pourra lui être d'aucun secours. Pour nous en assurer, essayons d'estimer le facteur temps par exemple dans le choix des paramètres du chasseur; si l'on admet que le passage d'un combat occupe l'ordinateur pendant une minute au moins, il faudra plusieurs heures pour une séquence de combats. Or, combien de séquences le constructeur doit-il passer pour obtenir une construction qui lui donne satisfaction?

Le constructeur doit améliorer les paramètres de l'avion. Il peut modifier une partie d'entre eux en se référant à sa propre expérience. Mais la modification d'un paramètre entraîne celle des autres. La situation devient « floue ». Plus le dispositif est compliqué et plus l'intuition manque de souffle. La question de savoir comment choisir les nouveaux paramètres pour que l'avion soit « meilleur » est très compliquée, même pour le plus génial des constructeurs. Finalement au lieu d'une recherche dirigée, on sera conduit à un tri qui demandera peut-être un millier d'itérations, c'est-à-dire un temps d'occupation de l'ordinateur qui dépasse toutes les limites du possible. La puissance de l'ordinateur n'est nullement incriminée ici. Cette situation est engendrée par la nature du problème et par sa complexité. C'est le point faible, le talon d'Achille de la simulation considérée comme instrument d'optimisation.

Le système de simulation est somme toute le banc d'essais des modèles. Le plan d'expériences revient bien moins cher que des essais en vol ou sur maquette grandeur nature. Mais ni lui ni les essais en vol ne peuvent servir à perfectionner de façon radicale le dispositif projeté. Le système de simulation est avant tout un instrument de mise en évidence des perfectionnements, même les plus insignifiants. Ce fait est manifestement très important en soi : avant de commencer les essais sur maquette, il faut réaliser un plan d'expériences, qui, on le sait, est considérablement moins onéreux. Et pourtant la mise en place d'un système de simulation ne résout pas le problème d'un coup de baguette magique.

Donc, pour tirer le meilleur parti du système de simulation et du

système automatique d'étude des projets il faut se rappeler ce qui a été dit *in limine*. Il ne faut pas perdre de vue par ailleurs que le constructeur en chef ne peut psychophysiologiquement maîtriser qu'une partie de l'information. D'où la nécessité d'instaurer un système de procédures qui permette au constructeur et tout d'abord au constructeur en chef de diriger la quête des paramètres optimaux sur la base de l'information qu'il est capable de maîtriser.

## § 2. Quelques variantes d'étude des projets

a) *Fonctionnelles auxiliaires, analyse parétienne.* Nous commencerons l'analyse des procédures d'étude automatique des projets par le plus haut niveau: celui du constructeur en chef. On a déjà signalé que le constructeur en chef pouvait raisonner par faible quantité de paramètres $F_j$, $j = 1, 2, \ldots, n$, des paramètres agrégats, c'est-à-dire des fonctions des paramètres $x^i$, $i = 1, \ldots, N$ $(n \ll N)$, de construction de l'avion. En réalité $n$ ne dépasse jamais la dizaine, $N$ est de l'ordre de plusieurs milliers.

De l'organisation et de l'utilisation des procédures non formelles il résulte que les caractéristiques agrégats par lesquelles raisonne l'expert sont toujours suffisamment particularisées. Est-ce à dire que les systèmes d'automatisation de l'étude des projets doivent être rigoureusement particularisés? Bien sûr que non. Les divers blocs du système, son organigramme, la structure des banques de données, le corps du logiciel doivent être banalisés. Il ne faut pas oublier par ailleurs le constructeur en chef qui a sa propre idée, ses propres appréciations de son modèle. Un système d'étude automatique des projets assez universel doit viser un bureau d'études bien défini. Ce qui revient à dire que le langage d'entrée doit être extensible et une certaine partie du logiciel, rédigée en fonction des exigences formulées par le constructeur en chef.

Il ne faut pas surestimer pour autant le rôle du facteur « particularisation ». De nombreux paramètres de construction sont universels. Pour un avion, par exemple, $F_1(x)$ est la plus grande vitesse, $F_2(x)$, la manœuvrabilité (le plus faible rayon de virage), $F_3(x)$, le plafond, etc. Mais d'autres paramètres spécifiques peuvent entrer en jeu selon la nature de l'avion projeté ou l'originalité du constructeur. La restructuration du logiciel sera une mince affaire, car ces calculs sont toujours accomplis par un bloc du système de simulation.

Il est important encore que le calcul des paramètres agrégats soit assez simple pour être à la portée du logiciel du système.

La simplicité de ce calcul est due au fait que les fonctionnelles auxiliaires ne dépendent, avec la précision nécessaire au constructeur en chef, que d'un nombre peu élevé de paramètres essentiels. D'où la possibilité de classer les paramètres de construction en deux grou-

pes :

$$x = (\hat{x},\ x^*),$$

où $\hat{x}$ est le vecteur des paramètres essentiels, sa dimension est peu élevée (de l'ordre de quelques dizaines) ; $x^*$, le vecteur de toutes les autres variables, sa dimension est de l'ordre de plusieurs milliers. Donc, la fonction $F_j(x)$ devient

$$F_j(x) = \hat{F}_j(\hat{x},\ \varepsilon,\ x^*), \tag{2.1}$$

où $\varepsilon$ est un petit paramètre tel qu'au niveau du constructeur en chef on peut toujours poser

$$\hat{F}_j(\hat{x},\ \varepsilon,\ x^*) \approx \hat{F}_j(\hat{x},\ 0). \tag{2.2}$$

REMARQUE. La possibilité de mettre $F_i(x)$ sous la forme (2.1) permet de développer et d'utiliser la technique de la théorie des perturbations exposée aux chapitres IV, V et VI.

De ce qui précède on peut encore tirer une importante conclusion : la résolution du problème d'optimisation

$$\hat{F}_j(x) \approx \hat{F}_j(\hat{x}, 0) \Rightarrow \max_{\hat{x} \in X} \tag{2.3}$$

est possible, et ce dans des délais raisonnables.

L'ensemble $X$ de l'expression (2.3) est l'ensemble des paramètres « possibles », c'est-à-dire tolérables par le niveau technologique actuel ainsi que par la conception de l'agencement de l'avion. A noter que la détermination de l'ensemble $X$, c'est-à-dire du système de contraintes, qui est l'une des étapes les plus difficiles de l'étude des projets, implique une très haute qualification de constructeur. C'est l'une des plus importantes procédures non formelles de l'automatisation de l'étude des projets.

Ainsi la première étape de la décomposition consiste à désigner un ensemble de fonctionnelles qui au regard du constructeur en chef caractérise assez complètement le dispositif afin de choisir parmi les variantes possibles celles qui seront analysées dans la suite. En d'autres termes, elle consiste à dégager des fonctionnelles réduisibles à une fonctionnelle de qualité qui ne pourra être décrite formellement. Appliquons-nous maintenant à développer des procédures qui nous permettront de choisir le dispositif en comparant des variantes qui appartiennent à l'ensemble de Pareto construit sur ces fonctionnelles. Le recensement de ces fonctionnelles est un acte non formel, mais ces fonctionnelles servent de base au développement d'un certain formalisme. L'étape suivante consiste à dégager les variables essentielles et à mettre les fonctionnelles $F_j$ sous la forme (2.1). Leur réduction à cette forme implique une étude profonde de leur structure.

Avant de passer à la description de l'étape suivante, c'est-à-dire l'organisation et l'utilisation des procédures d'optimisation de (2.3), procédures qui servent de base à la construction de l'ensemble de Pareto, attirons l'attention sur une circonstance que l'on rencontre inévitablement dans tout projet complexe.

L'étude des problèmes d'« optimisation des constructions » ou, plus exactement, l'étude des compromis possibles (c'est-à-dire d'optimisation parétienne) est toujours précédée du choix du « schéma architectural » de la future construction. Dans le cas d'un avion, c'est le choix de la composition. Les variantes de composition ne sont pas très nombreuses — monoplan à haute portance, schéma d'aile volante, schéma à trois moteurs, etc. —, elles forment un ensemble fini. Une fois qu'on a choisi une variante de composition, on commence à analyser et à sélectionner (on verra plus loin comment) les valeurs des paramètres. Si parmi l'ensemble des paramètres correspondant au schéma retenu (cet ensemble peut être continu ou discret) on ne trouve pas une collection satisfaisante, alors on prend une autre composition de la future composition, et ainsi de suite.

REMARQUES. 1. Ce phénomène, c'est-à-dire le choix d'une autre structure, fait partie d'une classe de phénomènes qui ont été étudiés pour la première fois par Poincaré. Cette théorie fut appelée théorie des bifurcations (voir à ce propos [18]). Un exemple classique de bifurcations est le changement de formes d'équilibre dans le problème d'Euler d'équilibre d'une barre. Les valeurs critiques du paramètre (la force de compression extérieure) correspondaient aux valeurs propres d'un problème aux limites, c'est-à-dire aux solutions d'un problème d'optimisation. Dans le cas étudié on change les formes architecturales lorsque le paramètre atteint sa valeur optimale. Ces dernières années les chercheurs se sont sérieusement penchés sur ces problèmes. La théorie générale des bifurcations est souvent appelée théorie des catastrophes (cf. par exemple [19]). A mon sens ce terme n'est pas très heureux, car il occulte la notion de bifurcation. De plus, l'introduction de ce nouveau terme semble faire peu de cas du passé de cette théorie et le lecteur non averti risque d'avoir une fausse idée de la nouveauté et de l'originalité de la nouvelle théorie.

2. La situation étudiée présente aussi un certain intérêt pour les recherches systémiques dans la mesure où elle donne une preuve éclatante de l'unité et de la contradiction de la forme et du fond : de la structure et de la fonction. C'est précisément sous cet angle que l'on peut interpréter la théorie des bifurcations de Poincaré. Les mathématiques ont jusqu'à maintenant étudié séparément les structures et les particularités de leur fonctionnement, c'est-à-dire d'un côté la géométrie contemporaine (et la topologie) et de l'autre l'analyse, la physique mathématique, etc. L'analyse et l'étude des projets de systèmes complexes impliquent, on le voit, un appareil susceptible de conjuguer ces deux principes. Je pense que la théorie de Poincaré peut servir de point de départ à l'élaboration de méthodes d'analyse numérique de telles situations.

Revenons maintenant aux procédures d'analyse parétienne, c'est-à-dire au choix des paramètres $\hat{x}$ réalisant le compromis :

$$\hat{F}_j(\hat{x}) \Rightarrow \max.$$

Une fois les fonctionnelles $\hat{F}_j$ définies, on doit exhiber une procédure pour décrire l'ensemble de Pareto construit sur l'espace de ces fonctionnelles. Nous aurons besoin pour cela des solutions des problèmes d'optimisation (2.3).

Désignons par $\hat{x}_i$ le vecteur $\hat{x}$ réalisant la solution du problème (2.3) pour la fonctionnelle $\hat{F}_i$. La résolution de ce problème nous donnera des nombres $\hat{F}_{i0} = \hat{F}_i(\hat{x}_i)$. Ces nombres caractérisent les possibilités limites des avions envisageables en fonction du niveau technologique.

L'avion qui possède les paramètres $\hat{x}_i = (\hat{x}_i^s)$ sera appelé avion du maximum de $i$. Il s'agira d'un avion qui aura par exemple le plus grand rayon d'action, la plus grande manœuvrabilité, le plus haut plafond, la plus grande charge utile, etc. En recherche opérationnelle on parle de schéma idéal ou de solution idéale. Les avions de maximum sont ces constructions idéales qui feront toujours partie du domaine de l'imaginaire.

Il est évident qu'aucun avion de maximum ne convient pour résoudre les problèmes qui ont régi sa construction. Mais un avion réel doit nécessairement combiner les principales qualités des avions de maximum. La construction réelle est une harmonisation des principales caractéristiques. L'avion réel doit être manœuvrable, atteindre une grande altitude, voler vite, être économique, etc. Cette harmonisation dépend bien sûr de la finalité de l'avion, de ses objectifs militaires et financiers, quant à la structure, la combinaison de ces caractéristiques, elles dépendent essentiellement de l'habileté du constructeur en chef.

On admettra donc que les critères $\hat{F}_i$ sont désignés par le constructeur en chef et les valeurs $\hat{F}_{i0}$, calculées par nous. Composons une nouvelle fonction objectif, ou, plus exactement, une famille de fonctions objectifs dépendant d'un paramètre vectoriel $\lambda$:

$$W(\hat{x}, \lambda) = \max_i \lambda^i \left(\frac{\hat{F}_{i0} - \hat{F}_i(\hat{x})}{\hat{F}_{i0}}\right). \tag{2.4}$$

Interprétons le sens de la formule (2.4). L'expression

$$[\hat{F}_{i0} - \hat{F}_i(\hat{x})]/\hat{F}_{i0}$$

prend toujours ses valeurs dans l'intervalle ] 0, 1 [. Elle donne le rapport de la qualité de l'avion réel à celle de l'avion du maximum de $i$ relativement à l'indice $\hat{F}_i$. Plus la valeur de cette expression est élevée et plus l'avion de paramètres $\hat{x} = \{\hat{x}^s\}$ s'écarte de l'avion du

maximum de $i$. L'expression

$$\max_i \frac{\hat{F}_{i0}-\hat{F}_i(\hat{x})}{\hat{F}_{i0}}$$

montre par rapport à quel paramètre et de combien l'avion réel s'écarte de l'avion du maximum de $i$.

Si l'on trouve un vecteur $\hat{x}_* \in X$ qui soit solution du problème

$$\max_i \frac{\hat{F}_{i0}-\hat{F}_i(\hat{x})}{\hat{F}_{i0}} \Rightarrow \min,$$

c'est que l'on aura trouvé un ensemble de paramètres $\hat{x}_*$ plus proches de ceux de l'avion de maximum sous réserve que tous les paramètres soient équivalents. Mais on sait que le rôle des paramètres varie avec la finalité de l'avion. Les nombres $\lambda^i > 0$ définissent la valeur relative de ces paramètres pour les objectifs posés. Plus $\lambda^i$ est élevé et plus la caractéristique $\hat{F}_i$ est importante pour nous. Le vecteur $\lambda = (\lambda^i)$ s'appelle vecteur des conceptions de l'avion. Il est naturel de le soumettre aux contraintes:

$$\lambda^i \geqslant 0, \quad i=1, \ldots, n, \quad \sum_{i=1}^{n} \lambda^i = 1. \tag{2.5}$$

La deuxième condition est une condition de normalisation ordinaire. Montrons maintenant quel usage on peut faire de la famille de fonctionnelles $W(\hat{x}, \lambda)$.

Rappelons d'abord ce qui a été dit sur l'optimum de Pareto. L'optimum de Pareto signifie que la meilleure construction que l'on se propose de choisir se trouve parmi les éléments de l'ensemble de Pareto construit sur les fonctionnelles $\hat{F}_i$. Cela signifie aussi que la fonctionnelle du problème est la réduction de toutes ces fonctionnelles. Les fonctionnelles $W(\hat{x}, \lambda)$ sont précisément obtenues par une telle réduction, puisque, en répertoriant les paramètres $\lambda$ vérifiant les conditions (2.5) et en cherchant pour eux les points $W(\hat{x}, \lambda) \Rightarrow \max_\lambda$, on obtient tous les points de l'ensemble de Pareto. Bien plus, on obtient des variantes dites semi-efficaces qui contiennent tous les points de l'ensemble de Pareto.

Donc, en conjecturant que les fonctionnelles $\hat{F}_i$ caractérisent assez complètement la construction, on est conduit au fait que cette construction doit être cherchée parmi les points de l'ensemble de Pareto construit sur les fonctionnelles $W(\hat{x}, \lambda)$.

Glissons provisoirement sur l'étude de l'optimum de Pareto et signalons que l'analyse des fonctionnelles $W(\hat{x}, \lambda)$ peut rendre de

grands services dans les cas où la fonctionnelle du problème, c'est-à-dire le critère qui régit le choix des paramètres de la construction, est connue mais difficile à calculer.

Supposons tout d'abord que le constructeur en chef sait formellement ce qu'il faut entendre par « optimisation de la construction ». Cela veut dire qu'il peut indiquer pour tout vecteur $\hat{x}$ une méthode formelle de détermination de la fonction $\Phi(\hat{x})$ dont le maximum pour $\hat{x} \in X$ nous donnera la meilleure construction. Nous avons vu que, s'agissant d'un avion de chasse, cette fonction était définie comme le pourcentage de combats gagnés au cours d'un plan d'expériences assez chargé. Etant donné que le modèle du duel est complètement formalisé, la valeur de la fonction $\Phi(\hat{x})$ se détermine à l'aide d'un algorithme. Mais dans le chapitre précédent on a signalé que la résolution directe du problème

$$\Phi(\hat{x}) \Rightarrow \max \qquad (2.6)$$

pouvait être pratiquement impossible à cause de la grande dimension du vecteur $\hat{x}$ et du fait que pour $\hat{x}$ fixe le calcul du critère $\Phi(\hat{x})$ demande beaucoup de temps machine.

Si l'on se sert de la fonctionnelle (2.4) on peut considérablement simplifier la procédure (2.6). Fixons le vecteur $\lambda = \hat{\lambda}$. La fonctionnelle $W(\hat{x}, \lambda)$ est composée d'un petit nombre de fonctionnelles $\hat{F}_i(\hat{x})$ (on rappelle que $i = 1, 2, \ldots, n$, où $n$ est un petit nombre, et que chaque $\hat{F}_i(\hat{x})$ se calcule facilement). Nous pouvons donc résoudre le problème

$$W(\hat{x}, \hat{\lambda}) \Rightarrow \max_{\hat{x} \in X}. \qquad (2.7)$$

La solution du problème (2.7) est un vecteur $\hat{x}_{\hat{\lambda}} = \hat{x}(\hat{\lambda})$, c'est-à-dire le vecteur des paramètres optimaux pour la conception $\lambda = \hat{\lambda}$ de l'avion.

On peut maintenant construire à l'aide du problème d'optimisation (2.6) une fonction

$$\hat{W}(\lambda) = W(\hat{x}(\lambda), \lambda), \qquad (2.8)$$

qui sera solution du problème

$$W(\hat{x}, \lambda) \Rightarrow \max_{x \in X} \qquad (2.9)$$

pour tous les $\lambda$ vérifiant les contraintes (2.5), c'est-à-dire pour toutes les conceptions de l'avion projeté.

Revenons maintenant au problème (2.6). Ce problème consiste à maximiser la fonctionnelle $\Phi$ sur l'ensemble $X$. Remplaçons-le par le problème d'approximation :

$$\Phi(\hat{x}) \Rightarrow \max_{\hat{x} \in \{\hat{x}(\lambda)\}} . \qquad (2.10)$$

Cette notation exprime que l'on cherche un vecteur $\hat{x}$ maximisant le critère $\Phi(\hat{x})$ sur l'ensemble des solutions du problème (2.9), c'est-à-dire sur l'ensemble de Pareto. L'ensemble $\{\hat{x}(\lambda)\}$ est sensiblement plus étroit que l'ensemble primitif et chacun de ses éléments satisfait visiblement la condition $\hat{x} \in X$.

On a ainsi remplacé le problème (2.6) par le problème (2.10) qui peut être mis sous la forme suivante:

$$\Phi(\hat{x}(\lambda)) = \hat{\Phi}(\lambda) \Rightarrow \max_{\lambda} . \qquad (2.11)$$

Le problème (2.11) est déjà plus simple que le problème primitif, car la dimension du paramètre $\lambda$ est bien plus petite que celle du vecteur $\hat{x}$. On rappelle que la dimension de $\lambda$ est égale au nombre des critères $\hat{F}_i$, nombre qui généralement ne dépasse pas la dizaine.

Au niveau du constructeur en chef la procédure générale d'étude se pose en les termes suivants.

1. Donner les fonctionnelles $\hat{F}_i(\hat{x})$ (acte essentiellement non formel).

2. Composer la fonctionnelle $W(\hat{x}, \lambda)$ (séquence de procédures rigoureuses).

3. Construire la fonction $\hat{x}(\lambda)$ : à cet effet introduire sur l'espace des $\lambda$ un réseau de nœuds $\lambda_k$ et résoudre le problème (2.7) pour chaque $\lambda = \lambda_k$.

4. Résoudre le problème (2.11) et déterminer la « valeur optimale » $\lambda = \lambda^*$.

5. Déterminer les paramètres $\hat{x}_* = \hat{x}(\lambda^*)$ et passer à l'étape suivante.

Ce système de procédures appelle deux remarques.

La première est que la détermination de l'extrémum de la fonction $\Phi(\hat{x})$ qui, en raison de la grande dimension du vecteur $\hat{x}$ et de la complexité du calcul de la fonction elle-même est impossible, a été remplacée par la recherche du maximum de $\Phi(\hat{x})$ sur un ensemble $\{\hat{x}(\lambda)\}$ plus étroit. La deuxième est que la structure de cet ensemble n'est pas le résultat uniquement de transformations formelles. Elle est basée sur le choix de fonctionnelles essentielles, c'est-à-dire

sur la constatation subjective du constructeur que l'avion optimal au sens du critère $\Phi(\hat{x})$ doit se trouver parmi ceux qui doivent satisfaire la condition

$$W(\hat{x}, \lambda) \Rightarrow \max$$

pour une valeur $\lambda$ qui évidemment est inconnue *a priori*.

Si donc la fonctionnelle $\Phi(\hat{x})$ est donnée sous une forme explicite, on peut en principe optimiser la construction sans recourir aux expertises. On aura seulement à étudier les conditions sous lesquelles la procédure d'optimisation en deux étapes fournit une approximation assez bonne de la construction optimale. En fait, on aurait pu regarder l'optimisation en deux étapes comme la recherche du maximum de la fonction $\Phi(\hat{x})$ sur l'ensemble de Pareto construit sur les fonctionnelles $\hat{F}_i(\hat{x})$. Pour construire cet ensemble il suffit de prendre une réduction quelconque des critères $\hat{F}_i$, par exemple la fonction $W(\hat{x}, \lambda)$ déjà construite, et trouver la fonction $\hat{x}_{\hat{\lambda}} = \hat{x}(\hat{\lambda})$ en résolvant le problème (2.7). L'ensemble des valeurs de $\hat{F}_i^*$ définies par les formules

$$\hat{F}_i^* = \hat{F}_i(\hat{x}(\hat{\lambda})) \tag{2.12}$$

forme une hypersurface dans l'espace des critères $\hat{F}_i$, définie paramétriquement par (2.12).

Signalons qu'au lieu de $W$ on aurait pu avec le même succès prendre la combinaison linéaire élémentaire

$$\Psi(\hat{x}, c) = \sum_i c_i \hat{F}_i(\hat{x}), \quad c_i > 0, \quad \sum_i c_i = 1.$$

Le choix de la fonctionnelle $W$ dans le projet de l'avion fut dicté exclusivement par des impératifs de suggestivité et de facilité de l'interprétation. Il est possible que d'autres fonctionnelles soient retenues dans l'étude des projets d'autres systèmes.

Dans la réalité les choses sont plus compliquées, car aucune fonctionnelle $\Phi(\hat{x})$ ne peut être composée sous une forme explicite. Le constructeur en chef ne peut comparer que les conceptions en se servant du système de simulation. Ce système entre en scène aussitôt que l'ensemble de Pareto est construit. Dans ces conditions, c'est-à-dire lorsque le plan d'expériences est peu chargé, le constructeur en chef doit choisir la conception qui lui semble le mieux correspondre à l'objectif initial. La tâche des mathématiciens et des projeteurs est de lui faciliter à l'extrême ce choix, de lui proposer la procédure de choix la plus simple. On pourrait par exemple procéder comme suit. On introduit dans l'espace des conceptions possibles un réseau de

nœuds $\lambda_k$, dont le pas devra être discuté avec le constructeur en chef. Le pas doit être plus fin dans le domaine de l'espace des conceptions où le succès est le plus probable.

REMARQUE. Signalons qu'à ce stade de l'analyse formelle le constructeur en chef est un membre actif sans lequel il est impossible d'instaurer un dialogue rationnel avec le système de simulation.

Pour chaque vecteur $\lambda_k$ on résout le problème (2.7) et on compose l'ensemble $\hat{x}_k = \hat{x}(\lambda_k)$. Cet ensemble peut être appelé ensemble des prototypes. Pour chacun de ces prototypes (et pour eux seuls) on organise un plan d'expériences dont on conviendra de désigner le résultat par $\Phi(\hat{x}_k)$. Ces résultats sont proposés au constructeur en chef, qui seul ou de concert avec le client commence le triage. Cette opération engendre un ensemble $S$ de l'espace des conceptions.

Ainsi, le schéma proposé donne lieu à un nombre peu élevé de plans d'expériences réalisables sur ordinateur.

A noter que les variantes sont traitées par lot contrairement à la méthode traditionnelle d'étude des projets. Après le premier triage, il reste encore plusieurs variantes en course. Ces variantes à chacune desquelles est associé un vecteur $\lambda_k$ sont soumises à une analyse plus détaillée. Si le constructeur et le client se fixent sur une conception $\lambda = \lambda^*$, on dira que le prototype est choisi. Si le constructeur en chef ne peut au terme de ces plans d'expériences se fixer sur un choix, c'est que, ou bien l'ensemble des critères $\hat{F}_i$ est trop étroit et il faudra alors l'élargir, ou bien le schéma de composition est inadéquat, ou bien la construction ne peut être réalisée avec les moyens existants.

Une fois le prototype choisi, on passe à l'étape suivante de l'étude du projet. Mais avant d'entrer dans le détail arrêtons-nous encore sur un cas particulier important.

b) *Cas d'existence d'une fonctionnelle dominante.* Jusqu'ici nous avons étudié le cas où le critère n'était pas formalisé, c'est-à-dire était une vision subjective de l'expert. Nous avons de même envisagé une situation dans laquelle on pouvait élaborer un système de procédures formelles permettant de calculer le critère. Mais ce calcul était si lourd qu'il était impossible de l'utiliser directement pour déterminer les paramètres optimaux de la construction. Dans les deux cas nous nous sommes servis d'un système de critères auxiliaires et dans les deux cas nous avons analysé l'ensemble de Pareto.

L'introduction de critères « essentiels » auxiliaires ne figurant pas directement dans la position du problème est un artifice génial qui donne toute sa dimension au dialogue homme-ordinateur. Cette idée est due à P. Krasnochtchokov ([46]). Nous avons ensuite utilisé une réduction de ces critères. Le choix de cette réduction est moins une question de principe que de commodité. Cette procédure n'est pas

universelle en dépit de son efficacité et de sa commodité : elle peut être profondément modifiée d'un cas à l'autre. Même l'analyse parétienne n'est pas toujours un élément obligatoire de l'étude des projets. En effet, les objets projetés sont assez souvent caractérisés par l'existence d'une fonctionnelle dominante et toute l'analyse de la construction est alors centrée sur l'étude des variantes au voisinage de l'optimum de cette fonctionnelle.

Supposons qu'un projet est caractérisé par les critères $J_0(x)$, $J_1(x), \ldots, J_N(x)$ et que le constructeur (ou le projeteur) tente de choisir les paramètres de la construction (ou du projet), c'est-à-dire le vecteur $x$, de manière à satisfaire les conditions suivantes :

$$J_i(x) \Rightarrow \min_x, \quad i = 0, 1, 2, \ldots, N \tag{2.13}$$

Le plus naturel semble-t-il est de se servir de l'analyse parétienne. Mais il n'est pas exclu que l'un des critères fasse l'objet d'une contrainte spéciale. Désignons par $x_0$ la solution du problème

$$J_0(x) \Rightarrow \min \tag{2.14}$$

Supposons que la contrainte en question implique que le vecteur $x$ soit tel que

$$J_0(x) \leqslant (1 + k)\, J_0(x_0), \tag{2.15}$$

où $0 < k \ll 1$. La fonctionnelle $J_0$ est généralement le coût du projet. La condition (2.15) exprime alors que quelles que soient les valeurs prises par les autres critères $J_1(x), J_2(x), \ldots, J_N(x)$, le coût du projet ne doit pas excéder le coût plancher $J_0(x_0)$ de $100\,k$ %. Ce cas qui est classique est caractéristique d'un grand nombre de projets économiques. Ainsi dans l'élaboration d'un projet d'exploitation d'une région pétrolifère (installations, forage, transport, traitement primaire, etc.), la caractéristique la plus importante est le prix de revient. Et la tâche première du projeteur est de déterminer le coût $J_0(x_0)$ du projet « idéal » *).

Une fois qu'on a déterminé le minimum de $J_0$ et le système de paramètres (le vecteur $x_0$) qui réalise ce projet idéal, on calcule les autres caractéristiques $J_1(x_0), J_2(x_0), \ldots, J_N(x_0)$ au voisinage de $x_0$. Ces données devront être présentées à l'expert qui en principe se montrera réservé. Donc, le projet le moins cher devra être mis au rebut. Il n'aura pas non plus l'aval du client pour les autres critères. Mais l'enveloppe allouée au projet nous oblige à serrer $J_0(x_0)$ de près. Donc, au voisinage du point $x_0$ il faudra construire un réseau de points auxquels sont associées les valeurs voisines de la fonctionnelle $J_0$. V. Khatchatourov qui a conçu un système automatique d'étude des projets de l'exploitation des gisements de pétrole et de

*) Dans ce cas généralement $k \leqslant 0{,}05$.

gaz propose de construire ce réseau en se basant sur la solution du problème (2.14) sous la condition

$$J_0(x) = (1 + k_l)\, J_0(x_0), \tag{2.16}$$

où les nombres $k_l \leqslant k$ sont selon les cas pris égaux à 0,01 ; 0,02 ; . . . On définit un réseau $x_s$ sur les hypersurfaces (2.16), on calcule les valeurs des critères $J_i(x_s)$, $i = 1, 2, \ldots, N$, qui sont présentées à l'expert, puis de l'ensemble des points $x_s$ on retient un sous-ensemble de variantes qui seront analysées ultérieurement.

Il est évident qu'on pourrait proposer d'autres méthodes de construction de l'ensemble des variantes au voisinage du point $x_0$. On pourrait en particulier envisager l'optique parétienne qui consiste à étudier la partie de l'ensemble de Pareto remplissant la condition (2.15).

Nous avons exposé une procédure de dialogue. Son principe présente beaucoup de parenté avec celui utilisé dans l'étude du projet d'avion sauf que l'ensemble de Pareto a été remplacé par un ensemble de variantes voisines. A noter que durant le dialogue peuvent apparaître de nouvelles contraintes (notamment de caractère non formel) et une nouvelle information. Tout ceci peut être facilement inclus dans le schéma proposé.

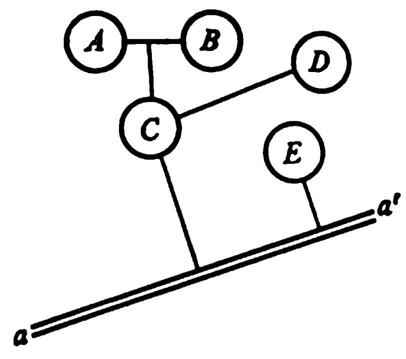



Fig. 8.1

c) *Autre exemple de décomposition.* La réalisation de la procédure décrite au numéro précédent se bute inévitablement à une difficulté inhérente à tout projet : la dimension du problème. Donc, avant d'entamer l'optimisation ou l'analyse d'un compromis (ou un dialogue) il faut simplifier le problème en le décomposant et élaborer un système d'algorithmes rapides. A cet effet nous avons introduit en a) les fonctionnelles et les variables essentielles qui ont permis de ramener des problèmes de plusieurs milliers de variables à des problèmes d'une ou de quelques dizaines de variables. Cette voie n'est pas unique, on pourrait en proposer d'autres, mais elles sont toutes liées d'une manière ou d'une autre à l'organisation d'une hiérarchie de problèmes dans le cadre du projet envisagé. Développons notre pensée sur l'exemple du projet d'un système d'exploitation d'une région pétrolifère.

Soit à élaborer le projet d'exploitation de gisements pétroliers $A$, $B$, $C$, $D$, $E$ (fig. 8.1) ([54, 39]). Quels problèmes doit résoudre le projeteur du plan général ? Tout d'abord il a un but bien précis : livrer la quantité de pétrole prévue dans le pipe-line collecteur $aa'$. Ce plan de livraison est un élément exogène. Il est imposé à chaque région en fonction des besoins nationaux en pétrole sous la forme d'une

fonction

$$Q = Q^*(t), \qquad t \in [t_0, T], \tag{2.17}$$

où $t_0$ est la date initiale d'extraction du pétrole, $T$, l'horizon du plan. L'intervalle $[t_0, T]$ couvre généralement une période de 20 à 30 ans pour l'exploitation des gisements de pétrole ou de gaz.

REMARQUE. La fonction (2.17) est construite au terme d'une analyse préliminaire de l'extraction du pétrole par les organismes de planification ou par l'Administration régionale. Elle peut être précisée et même modifiée au cours du projet en fonction de l'information recueillie. On admettra que la fonction $Q^*(t)$ qui représente la production globale de pétrole est donnée. Une fonctionnelle peut alors être rattachée à la production prévue par le plan. Elle peut par exemple être prise sous la forme

$$J_1 = \int_{t_0}^{T} (Q^* - Q)^2 \, dt. \tag{2.18}$$

A noter qu'un excédent de production est à double tranchant: si d'un côté il est utile, de l'autre il peut accélérer l'épuisement des gisements et rompre l'équilibre entre l'extraction, le transport et le traitement primaire du pétrole, c'est-à-dire au bout du compte se traduire par des pertes injustifiées.

Le projeteur doit définir les objectifs $Q_A^*$, $Q_B^*$, . . ., $Q_E^*$ des divers gisements, élaborer le projet de construction des pipe-lines reliant ces gisements au collecteur, établir l'ordonnancement des travaux, localiser les centres de stockage et de traitement primaire du pétrole, étudier un système de pompage de l'eau pour maintenir la pression en couche, déterminer l'emplacement des puits et le régime de forage, etc. Il faudra préparer une documentation complète sur tous les équipements nécessaires dont la nomenclature comprend des milliers d'articles. Cette multitude de variables doit être choisie de manière à réaliser non seulement le minimum de $J_1$, mais aussi celui de $J_0$ et de nombreux autres critères.

Il est évident que la rédaction du projet et le choix des paramètres impliquent une certaine hiérarchisation. L'étage supérieur est manifestement occupé par le plan général dans lequel chaque gisement est traité comme un objet autonome. Mais cette distinction du niveau supérieur n'a de sens que si chaque gisement est décrit par un nombre relativement peu élevé de paramètres. Mais quelle conduite adopter si le nombre de puits se chiffre en milliers dans certains gisements? Il est évident qu'il faudra envisager une forme spéciale d'agrégation des variables.

Le procédé d'agrégation est suggéré par la spécificité du problème. Puisque les puits sont en nombre très élevé, au lieu de les traiter comme des objets autonomes on les traitera comme des répartitions ou, ce qui est équivalent, on conviendra que leur nombre $N_i(t)$ sur un gisement est une fonction du temps continue et dérivable.

Le nombre sera alors régi par le système d'équations différentielles

$$\frac{dN_i}{dt} = f_i(u_i, t), \qquad (2.19)$$

où $N_i$ est le nombre de puits du gisement $i$, $u_i$ le vecteur ressources allouées au forage du gisement $i$ (nombre d'équipes de forage, dépenses d'argent ou de ressources matérielles). Soit $q_i(t)$, le débit d'un puits du gisement $i$. Cette quantité dépend de nombreux facteurs dont les plus importants sont le nombre $N_i$ de puits et la quantité de pétrole $Q_i$ déjà extraite. Donc, la loi de variation de $q_i$ est décrite par la relation paramétrique

$$q_i = g_i(N_i, Q_i). \qquad (2.20)$$

La quantité $q_i$ dépend encore de nombreux facteurs : du procédé d'exploitation, du niveau de la pression en couche maintenu par le pompage de l'eau, etc. Mais à l'échelon supérieur, où l'on n'a qu'une idée générale du projet, on admet que tous ces facteurs, dont nous sommes maîtres, sont définis « optimalement » (ou, plus exactement, sont imposés).

Ce procédé de paramétrisation est très répandu dans les études de projet et la planification à long terme. Ainsi par exemple dans le $\pi$-modèle de Yu. Ivanilov et A. Pétrov ([37]) où le paramètre essentiel est la date de démarrage des travaux, on admet que dès lors que les travaux ont débuté, les ressources nécessaires à leur poursuite seront allouées d'après une loi fixée *a priori*, bien qu'elles soient des commandes qui en principe peuvent être choisies d'une autre manière. La signification de l'imposition de certaines commandes est claire : ces commandes sont choisies à partir de la condition d'optimisation d'un critère auxiliaire, par exemple lors de la détermination des rythmes d'allocation des ressources pour l'achèvement de la construction de certaines entreprises du complexe industriel projeté. Ce critère sera, par exemple, le prix de revient de la construction de telle ou telle entreprise. Ceci est la conséquence d'une hypothèse plus ou moins évidente, à savoir que dans la plupart des cas toute entreprise mise en chantier doit être achevée dans les plus brefs délais, ce qui constitue un procédé de commande raisonnable du point de vue du complexe en général.

Revenons à l'équation (2.20). Les fonctions $g_i(N_i, Q_i)$ sont des fonctions convexes strictement décroissantes de l'ensemble de leurs variables : plus les réserves de pétrole s'épuisent (c'est-à-dire plus $Q_i$ est élevée), plus il devient difficile de les exploiter et moins chaque nouveau puits est rentable. Pour construire les fonctions $g_i(N_i, Q_i)$ il faut étudier le gisement et les conditions de son exploitation. Une fois les fonctions $g_i(N_i, Q_i)$ déterminées, on peut composer l'équation régissant les variations de $Q_i$ :

$$\frac{dQ_i}{dt} = q_i N_i = g_i(N_i, Q_i) N_i. \qquad (2.21)$$

Ainsi grâce aux hypothèses adoptées, on a décrit un gisement de façon élémentaire à l'aide des deux équations (2.19) et (2.21).

Si donc $k$ gisements sont exploités dans une région, on aura un système de $2k$ équations différentielles. Ces équations sont liées par les commandes $u_i$. Les commandes sont reliées par les relations

$$\sum_i u_i^j(t) = U^j(t) \tag{2.22}$$

ou

$$\sum_i \int_0^T u_i^s(t)\, dt = U^s. \tag{2.23}$$

La condition (2.22) est très importante, elle porte en général sur la main-d'œuvre : $U^j(t)$ est par exemple le nombre d'équipes de forage, c'est-à-dire le principal frein au développement de la région. Dans (2.23), la quantité $U^s$ représente la somme globale des matériaux dépensés à la construction ou la somme totale d'argent, etc.

Nous pouvons maintenant poser le problème d'optimisation : ventiller les ressources $u_i(t)$ de manière à minimiser la fonctionnelle (2.18). La résolution de ce problème nous permettra de déterminer les rythmes d'intensification de la production de tel ou tel gisement et notamment le début du forage des puits de chaque gisement, l'ordre de leur mise en service, etc.

Il est intéressant de noter que ces problèmes variationnels sont toujours dégénérés. Cette particularité importante exprime que leurs solutions forment une variété dans l'espace des commandes. La non-univocité de la solution des problèmes variationnels à l'échelon supérieur de l'étude du projet est une très importante propriété du schéma étudié. En effet, en étudiant les divers éléments du système, les projeteurs auront plus ou moins les coudées franches, puisque les fonctionnelles de la forme (2.18) peuvent être minimisées d'une infinité de façons. Au lieu de faire prendre aux commandes des valeurs « optimales » fixes, les projeteurs doivent seulement établir des relations entre elles.

Nous avons discuté pour le moment le problème de la ventilation des ressources pour la seule extraction du pétrole. Mais le problème primitif est bien plus compliqué : nous devons non seulement extraire le pétrole mais aussi l'acheminer jusqu'au collecteur. La quantité $Q(t)$ de l'expression (2.18) est la quantité de pétrole qui devra être livré aux consommateurs. Mais pour acheminer ce pétrole au collecteur il faut construire un réseau régional de conduites reliant les gisements au collecteur. Ces conduites forment un graphe orienté. Donc, le choix de la stratégie de construction sera limité non seulement par des contraintes sur les volumes des ressources, mais aussi par les liaisons logiques (des conditions de type ($\alpha$)) évoquées au chapitre précédent.

Il semble de prime abord que l'étude du projet du réseau de transport du pétrole et la ventilation des ressources pour l'extraction du pétrole peuvent être menées séparément. En effet, le réseau régional de conduites de pétrole doit être le moins cher possible, c'est-à-dire le plus court possible. Nous devons donc construire un graphe de sommets donnés dont la somme des longueurs des arêtes soit minimale. Lorsque ce problème aura été résolu, la forme du réseau d'acheminement du pétrole sera donnée et les conditions ($\alpha$) seront automatiquement proscrites. Mais en réalité la situation est plus compliquée. N'oublions pas en effet qu'il nous faut encore minimiser la fonctionnelle (2.18). Il nous faut conduire les travaux de manière à minimiser l'écart par rapport au plan $Q^*(t)$, or en essayant de construire le plus court graphe, on peut faire traîner la construction des conduites de pétrole à un point tel qu'il devienne impossible de commencer l'exploitation des gisements à temps. Donc, ces deux problèmes doivent être abordés simultanément.

Il est vrai que si l'on opte pour des procédures successives de calcul, on peut partir de la variante du graphe auquel correspond le réseau de conduites le moins coûteux (le plus court). Mais dans ce cas aussi la ventilation des ressources est un problème bien plus compliqué que celui qui a été traité, puisque parmi les travaux réalisés avec les mêmes ressources globales nous devons inclure les travaux nécessaires à la construction du réseau de conduites. De plus nous ne pourrons pas nous dédouaner entièrement des conditions ($\alpha$). En effet, l'une des plus importantes caractéristiques du réseau est le diamètre des conduites. Or ce diamètre dépend du débit du pétrole, c'est-à-dire des rythmes de production. Donc, même si le réseau d'acheminement du pétrole est donné, le problème reste assez compliqué, puisqu'il fait intervenir des éléments continus (l'utilisation des ressources est décrite par une fonction continue) et des éléments discrets (les paramètres des conduites de pétrole). Enfin, après avoir déterminé la structure du plus court graphe et résolu pour lui le problème de la ventilation optimale des ressources, on doit comparer cette solution avec celles obtenues pour de plus longs graphes.

REMARQUE. Les choses peuvent être simplifiées par le fait que les projets qui valent la peine d'être examinés ne peuvent différer du moins cher que de quelques pour cent de leur valeur.

Dans ce numéro nous avons étudié une méthode entièrement nouvelle de construction des algorithmes rapides. Si tous les exemples précédents étaient d'une manière ou d'une autre liés à l'analyse de la dépendance des propriétés du modèle par rapport aux paramètres et à l'élaboration des théories des perturbations correspondantes, dans l'exemple envisagé on ne trouve trace d'aucun paramètre explicite. Cette méthode, que l'on pourrait appeler méthode de continuation du problème, possède de nombreux analogues. En méca-

nique des fluides par exemple, le modèle de Boltzmann décrit le mouvement des particules de gaz dans un volume. Si les particules sont nombreuses, il n'y a aucun sens à considérer le mouvement individuel de chaque particule: on introduit alors de nouveaux paramètres agrégats: la densité, la vitesse, la température, etc. Une nouvelle description agrégée émerge: la mécanique des milieux continus. Cette description est certes moins précise dans la mesure où elle ne permet pas de localiser la position des molécules, mais par contre elle est notablement plus facile. C'est à peu de chose près ce que nous avons fait en décrivant les gisements en termes agrégats. Cette approche est trop grossière pour le projet technique d'exploitation d'un gisement, projet dans lequel nous devons décrire le fonctionnement de chaque puits. Mais elle est assez efficace pour fixer la date de démarrage de l'exploitation de tel ou tel gisement, répartir la main-d'œuvre entre les divers puits, déterminer le tracé du réseau, etc.

d) *Hiérarchie dans l'étude des projets et systèmes tampons.* Jusqu'à maintenant nous avons étudié uniquement des procédures réalisées à l'échelon supérieur. Ces procédures nous permettent d'obtenir une ébauche (ou avant-projet) du futur projet. L'importance de cette étape est indéniable. L'étape initiale du projet détermine l'avenir du système réel. A ce niveau les erreurs ne peuvent plus être corrigées.

En a) nous avons indiqué les possibilités offertes par l'analyse parétienne. Soulignons encore l'importance des propos tenus. Nous développons constamment le principe d'un dialogue, le système de procédures décrites ne vise qu'un seul objectif: éviter au constructeur de travailler dans le vide, d'analyser les variantes sans intérêt. Mais l'instrument décrit est d'une grande importance et pas seulement pour l'étude des projets de nouveaux systèmes. En effet, les procédures décrites nous permettent toujours d'apprécier les projets déjà élaborés et de les perfectionner au cas où ils ne seraient pas parétiens. Enfin, ces raisonnements peuvent être appliqués à un échelon différent de celui du constructeur en chef: au niveau de l'étude des composantes du système. Supposons donc qu'on a élaboré un avant-projet de l'avion. Cela veut dire qu'on a choisi un schéma de composition ainsi que les paramètres $\hat{x}$ dont dépendent les valeurs des fonctionnelles $\hat{F}_i$ (cf. formule (2.1)).

L'étape suivante consiste à étudier les projets des divers éléments de l'avion: fuselage, compartiments des moteurs, circuit électrique, etc. Chacun de ces éléments est complexe en soi et dépend d'une multitude de paramètres. La notion de « fuselage optimal » est aussi difficile à formaliser que celle d'« avion optimal ». Les principes qui ont régi l'élaboration de l'avant-projet de l'avion pourront donc être appliqués à l'avant-projet du fuselage, du moteur, etc. Il faudra seulement désigner de nouveaux critères auxiliaires qui en l'occurrence concerneront le poids, la fiabilité, le niveau technique, etc.

En évoquant les particularités des procédures de l'échelon supérieur, on a mis en évidence l'importance particulière de l'ensemble $X$ des contraintes. On peut en dire autant des contraintes portant sur les divers éléments de l'avion, sauf qu'aux containtes concernant les paramètres de construction, il faudra adjoindre les valeurs des paramètres $\hat{x}$ déterminés lors de l'élaboration de l'avant-projet de l'avion.

Supposons maintenant qu'on a projeté certains éléments de l'avion. A cette étape peuvent se présenter des difficultés de coordination des décisions prises au niveau du constructeur en chef et des constructeurs projetant les divers éléments de l'avion, soit que l'un de ces derniers, par exemple le constructeur du fuselage, sorte un modèle qui viole les normes fixées tant par son poids que par ses paramètres géométriques, soit qu'un autre, par exemple le constructeur du moteur, ne puisse pas réaliser la dépendance exigée de la puissance par rapport à l'altitude ou à la vitesse, etc. Donc, si l'on assemble ces divers éléments, on obtiendra un avant-projet très différent du modèle conçu par le constructeur en chef.

Une fois qu'on a fini de travailler sur l'avant-projet, c'est-à-dire qu'on a choisi les paramètres $\hat{x}$, on peut se le représenter plus en détail. Un programme susceptible de représenter cet avion en diverses projections permettrait au constructeur en chef de voir son modèle. Si à l'étape suivante où l'on entre davantage dans les détails du projet, on représente graphiquement l'avion, on verra encore apparaître d'autres dissonances. Il faudra envisager un nouveau système de calculs qu'on appellera système tampon. Ce système est constitué de nombreux programmes très spécifiques. Par exemple, nous pouvons maintenant calculer les fonctionnelles (2.1) avec plus de précision, car au second niveau nous connaissons la plupart des quantités $x^*$ et nous pouvons déterminer les valeurs des fonctionnelles

$$F_j(x) = \hat{F}_j(\hat{x}, \varepsilon x^*), \tag{2.24}$$

où $x^*$ sont, nous l'avons dit, des quantités connues et non plus des zéros comme nous l'avons admis en élaborant l'avant-projet. Donc les formules (2.24) et (2.3) nous donneront des valeurs différentes pour les critères $F_j(x)$.

Ceci nous suggère de développer une variante de la théorie des perturbations. Figeons par exemple la valeur trouvée $x_1^*$ des paramètres $x^*$, considérons le système de fonctionnelles $F_{j1} = \hat{F}_j(\hat{x}, \varepsilon \hat{x}_1)$, qui dépendent encore uniquement de $\hat{x}$, et reprenons les calculs effectués pour l'avant-projet. Nous obtenons un nouvel ensemble parétien et par suite de nouvelles versions de prototypes. Cet ensemble plus précis tient compte des valeurs des paramètres $x^*$ qui dans la première analyse ont été purement proscrits.

Le système tampon permet de préciser la conception. On pourrait envisager d'autres systèmes de reprise des calculs. Supposons par exemple que le constructeur du fuselage n'a pas réussi à respecter le système de contraintes ou que les ailes ne satisfont pas les critères de résistance à cause de leur géométrie. Donc, dans ce cas les paramètres $\hat{x}$ seront différents des paramètres théoriques: $\hat{x}_1 \neq \hat{x}$. Ce qui signifie que les performances de l'avion seront altérées. La première chose à faire est d'évaluer les pertes et de recalculer toutes les caractéristiques $\hat{F}_j(\hat{x}_1)$ en tenant compte des corrections apportées aux échelons inférieurs. Nous évaluons ainsi l'adéquation du changement des paramètres $\hat{x}_1$. Si ce changement retentit sur les performances dans les limites autorisées, on peut passer au niveau suivant de l'étude du projet. Sinon cette reprise des calculs n'est qu'une étape du dialogue.

Donc, le système tampon remplit une fonction multiple. Il est à la fois instrument de précision des paramètres du dispositif construit et instrument de dialogue entre les projeteurs. La structure du système tampon est fortement individualisée et il est par conséquent très difficile d'émettre des recommandations générales pour son élaboration. Une chose est sûre toutefois: étant instrument de dialogue, le système tampon devra nécessairement comporter un bon système de programmes permettant d'instaurer facilement le dialogue avec l'ordinateur.

Notre propos se généralise à tout autre dispositif technique complexe: vaisseau, fusée, chaîne de production, entreprise chimique, etc. Les structures des algorithmes et critères seront certes différents, mais l'approche, la décomposition du problème, l'analyse parétienne et les méthodes d'instauration du dialogue par le système de simulation seront les mêmes.

Dans ce chapitre nous avons traité un autre problème: l'exploitation d'une région pétrolifère. Là aussi nous avons retrouvé les mêmes écueils. Après avoir déterminé au niveau supérieur la structure du système et ses principaux paramètres — les dates de forage et les rythmes de production $Q_i(t)$ des divers gisements, les flots de ressources — on peut passer à l'étude des projets d'exploitation de chaque gisement. Les paramètres définis à l'échelon supérieur serviront de points de départ (ou de contraintes).

**REMARQUE.** Malheureusement les ingénieurs accordent beaucoup d'attention aux objets du bas de l'échelle et pas assez à la région dans son ensemble. Dans le cas de la production pétrolière, le réseau d'acheminement à l'intérieur du gisement, le régime de forage et de maintien de la pression en couche, etc., ont été étudiés avec la plus grande minutie. Par contre l'étude globale du complexe et l'analyse des interactions régionales, des méthodes et des cadences de production du pétrole et du gaz avec le développement économique national n'ont pas fait l'objet d'un profond traitement mathématique. Si dans l'étude des projets de systèmes techniques de type avion, il est traditionnel de mener

l'analyse de haut en bas, c'est-à-dire du système global aux composants, dans l'étude des complexes économiques c'est la démarche inverse qui prévaut, c'est-à-dire qu'on fignole les composants et on passe la main sur le modèle complet. Nous avons déjà signalé l'inadmissibilité d'une telle approche.

* * *

Sur ces remarques nous fermons le chapitre consacré à l'automatisation de l'étude des projets, l'une des plus importantes applications de l'analyse des systèmes. Nous avons envisagé deux exemples d'étude des projets qui dans une certaine mesure sont extrêmes *). Néanmoins les études des projets des systèmes d'automatisation présentent de nombreuses affinités qui nous permettent de traiter l'automatisation comme une sorte de dialogue spécialement conçu. L'instauration du dialogue homme — machine est primordiale dans l'automatisation de l'étude des projets. Notre objectif numéro un a été de montrer que le dialogue est un algorithme original dont la confection requiert une grande qualification et que c'est la pierre angulaire de l'automatisation de l'étude des projets.

---

*) Nous avons envisagé des projets de systèmes réellement compliqués nécessitant des recherches interdisciplinaires et l'utilisation d'une information très variée. Les projets du matériel radio-électronique, de plaquettes à circuit imprimé, etc., sont à l'étude actuellement. Malgré la complexité des algorithmes, ces dispositifs sont simples, car leur fonctionnement peut être décrit dans le cadre de modèles simples et stéréotypés. L'étude automatique des projets de tels systèmes est notablement plus simple. Nous avons glissé sur ces systèmes bien qu'une grande partie de notre exposé puisse être appliquée à l'étude de leurs projets.

# BIBLIOGRAPHIE

## Principale


1. Bellman R., Kalaba R.— *Quasilinearization and nonlinear boundary-value problems*. N. Y., 1965.
2. Ermoliev Yu.— *Méthodes de programmation stochastique*. Moscou, « Naouka ». 1976 (en russe).
3. Hermeyer Yu.— *Introduction à la théorie de la recherche opérationnelle*. Moscou, « Naouka », 1971 (en russe).
4. Hermeyer Yu. — *Jeux à intérêts non opposés*. Moscou, « Naouka» , 1976 (en russe).
5. Moïsséev N.— *Eléments de théorie des systèmes optimaux*. Moscou, « Naouka », 1975 (en russe).
6. Moïsséev N.— *Les mathématiques effectuent les expériences*. Moscou, « Naouka », 1979 (en russe).
7. Moïsséev N.— *Méthodes asymptotiques de mécanique non linéaire*. Moscou, « Naouka », 1981 (en russe).
8. Moïsséev N., Ivanilov Yu., Stoliarova H.— *Méthodes d'optimisation*. Moscou, «Naouka », 1978 (en russe).
9. Pontriaguine L., Boltianski V., Gamkrélidzé R., Michtchenko E.— *Théorie mathématique des processus optimaux*. Traduction française. Editions « Mir », 1974.
10. Pougatchev V. — *Eléments de commande automatique*. Moscou, « Naouka », 1968 (en russe).
11. Rastriguine L.— *Systèmes de commande extrémale*. Moscou, « Naouka », 1974 (en russe).
12. Stronguine R.— *Méthodes numériques dans les problèmes à plusieurs extrémums*. Moscou, « Naouka », 1978 (en russe).
13. Ventsel H.— *Introduction à la recherche opérationnelle*. Moscou, « Sov. radio », 1964 (en russe).
14. Volterra V.— *Leçons sur la théorie mathématique de la lutte pour la vie*. Paris, Gauthier-Villars, 1931.
15. Wilde D. J. — *Optimum seeking methods*. N. J., Englewood Cliffs, 1964.


## Complémentaire


16. Abramov A.— Sur le transfert des conditions aux limites pour les systèmes d'équations différentielles ordinaires linéaires. *JVM i MF*, 1961, n° 3 (en russe).
17. Akoulenko L., Tchernoousko F.— Méthode de moyennisation dans les problèmes de commande optimale. *JVM i MF*, 1975, 15, n° 5 (en russe).
18. Andronov A., Léontovitch E., Gordon I., Mayer A.— *Théorie des bifurcations de systèmes dynamiques dans le plan*. Moscou, « Naouka », 1967 (en russe).

19. Arnold V.— Théorie des catastrophes. *Priroda*, 1979, nº 10 (en russe).
20. Beïko M., Beïko I. — Sur une nouvelle approche de la résolution des problèmes aux limites non linéaires. *Ukr. matem. journal*, 1978, 20, nº 6 (en russe).
21. Bertalanffy L.— *Theoretische Biologie*. Berlin, 1932.
22. Bogdanov A.— *Théorie de l'organisation ou Tectologie*. Moscou, 1913 (en russe).
23. Chatrovski L.— Sur une méthode numérique de résolution des problèmes de commande optimale. *JVM i MF*, 1962, nº 2 (en russe).
24. Danilov K., Schwarz S.— Sur les macrosystèmes biologiques. *In*: *Science et humanité*. Moscou, « Znanié », 1975 (en russe).
25. Erouguine N.— *Recueil de textes sur le cours général d'équations différentielles*. Minsk, « Naouka i tekhnika », 1979 (en russe).
26. Erechko F., Zlobine A.— Optimisation d'une forme linéaire sur un ensemble efficace. *In*: *Méthodes numériques de programmation non linéaire*. Kharkov, 1976 (Troudy II vsésoyouznogo séminara) (en russe).
27. Evtouchenko Yu.— Calcul aproché de commande optimale. *PMM*, 1970, 34, nº 1 (en russe).
28. Evtushenko Ju. G.— Approximate calculation of optimal control by averaging method. *Lecture notes in mathematics*, 112, Springer-Verlag, 1970.
29. Gantmaher F.— *Théorie des matrices*. Moscou, « Naouka », 1967 (en russe).
30. Glouchkov V.— Prédiction sur la base des expertises. *Kibernétika*, 1969, nº 2 (en russe).
31. Goloubev V.— *Cours de théorie analytique des équations différentielles*. Moscou, 1950 (en russe).
32. Gorélik V.— Systèmes hiérarchiques à structure rhomboïdale. *In* : Thézissy dokl. III Vses. konf. po issled. opératsiï, Gorki, 1978 (en russe).
33. Gratchev N., Evtouchenko Yu.— Application de la méthode des perturbations singulières à la résolution des problèmes de minimax. *DAN SSSR*, 1977, 233, nº 3 (en russe).
34. Grozdovski G., Ivanov Yu., Tokarev V. — *Mécanique du vol spatial à faible propulsion*. Moscou, « Naouka », 1966 (en russe).
35. Hermeyer Yu., Vatel I. — Jeux à vecteur des intérêts hiérarchique. *Tekh. kibernétika*, 1974, nº 3 (en russe).
36. Isaac R.— *Differential Games*, N. Y., Wiley, 1965.
37. Ivanilov V., Pétrov A.— Modèle dynamique d'extension et de restructuration de la production (π-modèle). *In: La cybernétique au service du communisme*, vyp. 6, Moscou, Energuia », 1971 (en russe).
38. Jouravlev Yu.— Algèbres correctes sur des ensembles d'algorithmes (euristiques) incorrects, I. *Kibernétika*, 1977, nº 4 (en russe).
39. Khatchatourov V.— Construction d'un système de simulation pour la planification du développement d'une nouvelle région pétrolifère. *In*: Travaux de la conférence internationale « Simulation des processus économiques ». Moscou, VTs. AN SSSR, 1975 (en russe).
40. Kolmanovski V., Moïsséev N., Tchernoousko F.— Sur une méthode de commande de systèmes stochastiques. *Tekh. kibernétika*, 1979, nº 3 (en russe).
41. Kononenko A.— Analyse théorico-ludique d'un système de commande hiérarchique à deux niveaux. *JVM i MF*, 1974, nº 5 (en russe).
42. Kononenko A.— Théorie de la commande et structures hiérarchiques. *In*: *Problèmes de gestion de systèmes économiques dirigés*. Novossibirsk, « Naouka », 1975 (en russe).
43. Kononenko A.— Sur les stratégies d'équilibre dans les jeux différentiels non antagonistes. *DAN SSSR*, 1976, 231, nº 2 (en russe).
44. Kononenko A.— Sur les conflits à coups multiples avec échange d'information. *JVM i MF*, 1977, nº 4 (en russe).
45. Koukouchkine N.— Jeux non coopératifs à trois personnes à structure hiérarchique fixe. *JVM i MF*, 1979, nº 4 (en russe).

46. Krasnochtchekov P., Morozov V., Fédorov V.— Décomposition dans les problèmes d'étude des projets. *Tekh. kibernétika*, 1979, nº 2 (en russe).
47. Krassovski N.— Théorie de la commande du mouvement. Moscou, « Naouka », 1968 (en russe).
48. Krylov I., Tchernoousko F.— Sur la méthode des approximations successives dans la résolution des problèmes de commande optimale. *JVM i MF*, 1962, nº 6 (en russe).
49. Lébédev V.— Calcul du mouvement d'un engin spatial à faible poussée. Série: Méthodes mathématiques dans la dynamique des engins spatiaux. Moscou, VTs AN SSSR, 1968, nº 5 (en russe).
50. Lioubouchkine A.— Convergence de la méthode du petit paramètre pour les systèmes optimaux faiblement gouvernés. *PMM*, 1978, 42, nº 3 (en russe).
51. Lotov A.— Sur la notion d'ensembles permis généralisés et leur construction pour des systèmes commandés linéaires. *DAN SSSR*, 1980, 250, nº 5 (en russe).
52. Méthodes de l'analyse des systèmes dans les problèmes de distribution des ressources hydrauliques, t. 1. IISA, Vienne, 1974.
53. Mikhalévitch V.— Algorithmes séquentiels d'optimisation et leur application. *Kibernétika*, 1965, nº 12 (en russe).
54 Moïsséev N., Khatchatourov V. — Automatisation de l'étude des projets d'exploitation de nouvelles régions pétrolifères. *In*: Actes du Symposium soviéto-finlandais « Systèmes automatiques d'étude des projets ». Moscou, VTs AN SSSR, 1977.
55. Molodtsov D.— Commande adaptative dans les jeux répétitifs. *JVM i MF*, 1978, nº 1 (en russe).
56. Okhotsimski D.— Sur la théorie du mouvement des missiles. *PMM*, 1946, 10, nº 2 (en russe).
57. Pantell R. H.— *Techniques of environmental systems analysis*. John Wiley & Sons, Inc. N. Y., 1976.
58. Ponomarev V.— Méthode d'optimisation sequentielle dans les problèmes de commande. *Tekh. kibernétika*, 1967, nº 3 (en russe).
59. Ponomarev V., Ptouchkine A.— Optimisation séquentielle d'un système de commande discret. *Tekh. kibernétika*, 1967, nº 3 (en russe).
60. Pougatchev V.— Problème général du mouvement dans l'air d'un projectile d'artillerie tournant. *In*: Travaux de l'Académie de l'air Joukovski, 1940. nº 70 (en russe).
61. Polya G.— *Mathematics and plausibles reasoning*. Princeton, N.Y., Princeton Univ. Press, 1954.
62. *Problème de planification et de gestion* (Pospélov G., Ven V., Solodov V., Chafranski V., Erlikh A.). Moscou, « Naouka », 1980 (en russe).
63. Schmalhausen I.— Bases du darwinisme. Moscou, « Naouka », 1960 (en russe).
64. De Sparre.— *Sur le mouvement des projectiles oblongs*. Paris, Imprimerie nationale, 1893.
65. Tamarkine Ya.— Sur certains problèmes généraux de théorie des équations différentielles et sur le développement des fonctions en série. Pétrograd, 1917 (en russe).
66. Tchernoousko F.— Quelques problèmes de commande optimale à petit paramètre. PMM, 1968, 32, nº 1 (en russe).
67. Tchétaev F.— *Stabilité du mouvement. Travaux de mécanique analytique*. Moscou. Izd-vo AN SSSR, 1962 (en russe).
68. Tikhonov A.— Système d'équations différentielles contenant de petits paramètres en les dérivées. *Matem. zbornik*, 1952, 31 (73), nº 3 (en russe).
69. Vassiliéva A., Boutouzov N.— Représentation asymptotique des solutions des équations perturbées singulières. Moscou, « Naouka », 1973 (en russe).

70. Vatel I.— Sur les modèles mathématiques de stimulation en économie. *In*: *Problèmes de gestion des systèmes économiques dirigés.* Novossibirsk, « Naouka », 1975 (en russe).
71. Vatel I., Dranev Ya.— Sur une classe de jeux répétitifs à information incomplète dans un système économique à deux niveaux. *In*: Travaux de la conférence internationale « Simulation des processus économiques ». Moscou, VTs AN SSSR, 1975 (en russe).
72. Vatel I., Erechko F.— *Les mathématiques des conflits et de la coopération.* Moscou, « Znanié », 1973 (en russe).
73. Ventsel D., Chapiro Ya. — *Balistique extérieure*, t. II, Moscou, « Oboronguiz », 1939 (en russe).

# INDEX TERMINOLOGIQUE

